Universidade de Aveiro
2010

Departamento de Electrónica, Telecomunicações e Informática

Jair José Lopes Delgado

Sistema de Informação de Apoio à Detecção de Perdas de Energia Eléctrica – O Caso da Electra

Universidade de Aveiro
2010

Departamento de Electrónica, Telecomunicações e Informática

Jair José Lopes Delgado

Sistema de Informação de Apoio à Detecção de Perdas de Energia Eléctrica – O Caso da Electra

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Electrónica e Telecomunicações vertente Sistema de Informação, realizada sob a orientação cientifica dos Professores Doutor José Manuel Matos Moreira, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro e do Professor Doutor José Maria Amaral Fernandes, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro.

Com o apoio da Cooperação Portuguesa

COOPERAÇÃO PORTUGUESA

Dedico este trabalho à minha família.

O Júri

Presidente	Prof. André Ventura Zúquete Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro
Arguente	Prof. Carlos Alberto Baptista de Sousa Pinto Professor Auxiliar do Departamento de Sistemas de Informação da Escola de Engenharia da Universidade do Minho.
Vogais	
Orientador	Prof. José Manuel Matos Moreira Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro
Co-orientador	Prof. José Maria Amaral Fernandes Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Agradecimentos

Gostaria imenso de expressar os meus agradecimentos a minha família pelo apoio moral e incondicional no incentivo e razão de todos os esforços, aos meus colegas e amigos pela coragem e por sempre acreditarem nas minhas capacidades e principalmente aos meus orientadores pela colaboração preciosa, com opiniões e sugestões ao longo da realização do trabalho.

Um obrigado especial aos profissionais da empresa, Electra, que se disponibilizaram a responder às questões das entrevistas.

Muito Obrigado

Palavras-Chave

Data Warehousing, *OLAP*, *Data Mining*, Árvores de Decisão, Redes de Distribuição de Energia Eléctrica, Detecção de Fraudes, Perdas de Energia Eléctrica.

Resumo

A realidade mundial é preocupante no que diz respeito ao aumento de ocorrências de perdas e fraudes em redes de distribuição de energia eléctrica. Em Cabo Verde, mas precisamente na Cidade da Praia a realidade é ainda mais preocupante devido ao número de ocorrências e a gravidade dos mesmos.

Propõe-se um trabalho de investigação sobre perdas e fraudes de energia eléctrica baseado na análise dos dados relativos aos registos dos clientes na Base de Dados da Electra (Cabo Verde), com o intuito de nortear as tomadas de decisões de gestão estratégica no que diz respeito às políticas de controlo e prevenção de perdas e fraudes de energia eléctrica.

O trabalho baseia-se na recolha e selecção de dados a organizar numa *Data Warehouse* para depois aplicar as tecnologias *OLAP* para a identificação de perdas nos Postos de Transformação e zonas geográficas da Cidade da Praia em Cabo Verde e posteriormente identificar possíveis fraudes de energia eléctrica nos clientes finais utilizando *Data Mining*.

Os resultados principais consistiram na identificação de situações de perdas de energia eléctrica nos Postos de Transformação, a identificação de áreas críticas seleccionadas para inspecção dos seus clientes finais e a detecção de padrões de anomalias associadas ao perfil dos clientes.

Keywords *Data Warehousing, OLAP, Data Mining, Decision trees, Electric Power Distribution Networks, Fraud Detection*

Abstract

This work focuses on the study of losses and frauds' detection on electric power distribution networks, based on the analyses of the customers' records in the Electra (Cabo Verde) database. The aim of this research study is to guide the strategic management decisions, related with the policies for control and prevention of losses and frauds in the electric power distribution network.

This work includes data collection, transformation and organization in a Data Warehouse and subsequent application of OLAP technologies identify the losses in the transformation posts and geographic regions, followed by the identification of possible frauds in by the final costumers' using the Data Mining techniques.

The main results of this work are: the analyses and discovery of the loss of power in the transformation posts, the identification of critical areas for inspection of the final consumers and the detection of anomalies based on the profile of the client.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

Acrónimos e terminologia técnica

Alta Tensão – Tensão entre fases cujo valor eficaz é superior a 60 Kv.

Baixa Tensão – Levam a energia eléctrica desde os Postos de Transformação, ao longo das ruas e caminhos até aos locais onde é consumida em Baixa tensão (230 V entre fase e neutro e 400 V entre fases).

Cliente – Pessoa singular ou colectiva que, através de um contrato de fornecimento ou de um Acordo De Acesso às Redes, compra energia eléctrica para consumo próprio.

Consumo de Energia – Quantidade de energia eléctrica utilizada por um consumidor, que é oferecida e medida pela distribuidora do sistema eléctrico num determinado período. A grandeza que a define é o kWh (Quilowatt-hora), e sua unidade base é o Watt.

DM – *Data Mining* – é o processo não trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis.

DW - *Data Warehouse* - é uma base de dados de grande porte, resultado da junção de vários sistemas de base de dados ou técnicas que aplicadas em conjunto, servirão para gerar um sistema de dados, que tem por objectivo fornecer suporte na criação de relatórios, visando gerar informações que auxiliam nas tomadas de decisões de uma empresa.

Energia eléctrica - Grandeza escalar que caracteriza a aptidão de um sistema físico para realizar trabalho.

ETL – Extracção, transformação e carga – é uma das fases mais críticas e árduas da modelação de um *Data Warehouse*, pois nela ocorre padronização, modelação, limpeza, integração e transformação de grandes volumes de dados.

Indicadores - São variáveis perfeitamente identificáveis, utilizadas para caracterizar (quantificar ou qualificar) os objectivos, metas ou resultados.

KDD – *Knowledge Discovery in Databases* - podem ser vistos como processos de descoberta de novas correlações, padrões e tendências significativas por meio da análise minuciosa de grandes conjuntos de dados.

Média Tensão - São aquelas cuja tensão nominal é inferior a 60 kV. As tensões mais

comuns são 10,15 e 30 kV. Estas linhas ligam as subestações aos Postos de Transformação ou ligam diferentes Postos de Seccionamento/Transformação entre si.

OLAP - (*On-Line Analytical Processing*) ferramenta de acesso ao *Data Warehouse* que representa um conjunto de processos projectados para suportar análise e consultas. Os sistemas *OLAP* ajudam analistas e executivos a sintetizarem informações sobre a empresa, através de comparações, visões personalizadas, análise histórica e projecção de dados em vários cenários.

Perdas – Diferença entre a energia que entra num sistema eléctrico e a energia que sai desse sistema eléctrico, no mesmo intervalo de tempo.

Portinholas - Genericamente estas portinholas contêm entradas onde chegam e de onde partem a tensão de baixa tensão, e acessórios de protecção. Estas portinholas estão protegidas por fusíveis entre outros dispositivos de protecção.

PT – Instalação eléctrica destinada a transformação de energia de um nível de tensão elevado para o nível de tensão de utilização: 15kV para 380 a 250kV.

SBC – Sistema Baseado em Conhecimento - pode ser definido como sendo um sistema computacional que representa e utiliza conhecimento para resolução de problemas.

SGBD – Sistema de Gestão de Base de Dados – é um conjunto de programas responsáveis pela gestão da base de dados.

Subestações Entende-se a Instalação eléctrica destinada à transformação de energia de um nível de tensão elevado, para outro nível de tensão mais baixo e vice-versa.

TIC – Tecnologia de Informação e Comunicação – Pode ser definido como um conjunto de hardware e software usado para adquirir, transmitir, processar e expandir informação, bem como as metodologias de planeamento e desenvolvimento de sistemas de informação.

Transformador – Têm a função de reduzir a Média Tensão para a Baixa Tensão utilizável pelo consumidor final doméstico, comercial ou pequeno industrial.

Weka – Ferramenta de *Data Mining* que procede à análise computacional e estatística dos dados fornecidos recorrendo a técnicas de *Data Mining* tentando, indutivamente, a partir dos padrões encontrados gerar hipóteses para soluções.

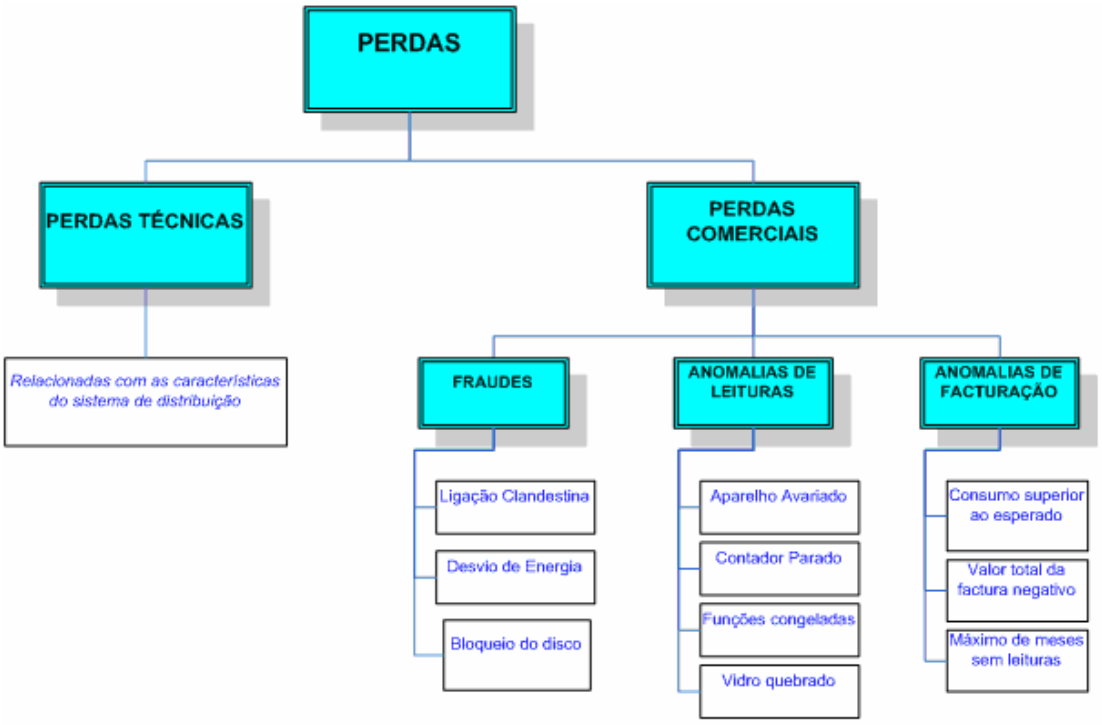
1. Introdução

1.1. Contextualização

A única empresa de produção e distribuição de energia em Cabo Verde, Electra, vem enfrentando grandes desafios. Atropelada pela crise energética pela qual o país passa e pelas novas regras de gestão mais rigorosa, ela está actualmente obrigada a buscar novos meios para optimizar a sua operação e maximizar a qualidade dos serviços de forma a garantir a sua rentabilidade e sobrevivência no mercado e possa prestar um serviço essencial para a sociedade.

A distribuição de energia eléctrica implica perdas que podem ser técnicas e comerciais. As perdas técnicas podem ocorrer naturalmente no processo de distribuição de energia, ou seja, relacionadas com as características do próprio sistema de distribuição. As perdas comerciais estão relacionadas directamente com as fraudes (ligações clandestinas), falhas na medição, erros de leituras e facturação, que fazem o desvio de energia eléctrica da rede de distribuição directamente para as instalações do consumidor, sem passagem pelo contador de energia (Eller, 2003).

O conceito de perdas, no âmbito da energia eléctrica, pode ser definido pela diferença entre os valores fornecidos e medidos de energia eléctrica numa região num dado intervalo de tempo, enquanto que fraudes podem ser actos praticados (bloqueio do disco do contador, ligações clandestinas, desvios de energia) intencionalmente para lesar terceiros ou são violações de obrigação usando procedimentos aparentemente ilícitos.



Esquema 1 - Conceito de Perdas

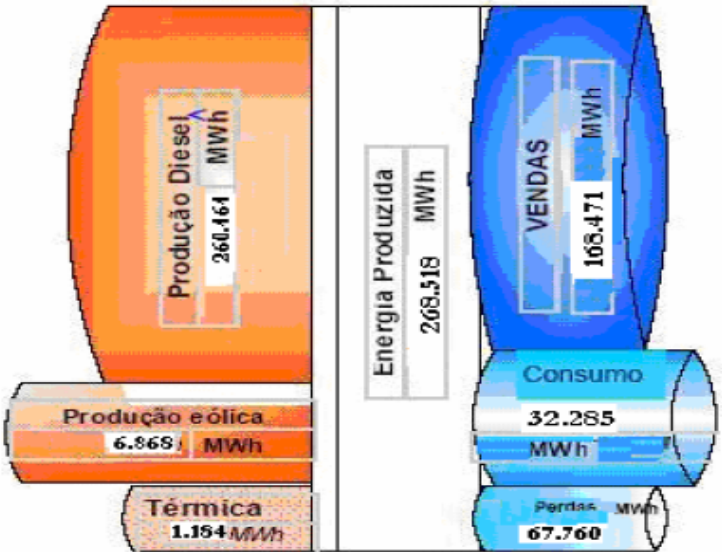
Neste momento, a Electra está a iniciar a integração das Tecnologias de Informação e Comunicação (TIC) no negócio ao nível da produção e da gestão dos clientes de forma a diminuir o risco de perdas, seja através da diminuição de clientes ou de perdas não identificadas ao nível do consumo. A possibilidade de poder contar com ferramentas automatizadas detalhadas sobre o perfil dos seus clientes e com sistemas de suporte ao processamento da informação pode permitir a detecção de perdas comerciais na rede eléctrica, o que se pode traduzir em benefícios para a empresa. A ocorrência de fraudes pode ser considerada, sem dúvida, uma das principais causas de perda de receitas em diversas áreas de negócios (Delaiba et al., 2004).

O aumento das perdas comerciais em energia eléctrica em todo o país, traduz-se na queda de receita e perda de energia eléctrica. Assim, buscar alternativas para minimizar as perdas tem uma relevância estratégica para a Electra. No entanto, o problema de fraude de energia eléctrica tem-se mostrado de difícil solução devido à variedade de maneiras de subtrair ilicitamente a energia e ao elevado custo de equipamentos capazes de detectar ocorrências de irregularidades.

1.2. Motivação do Trabalho

Para a Electra é de importância vital do ponto de vista do negócio fazer desaparecer (idealmente) ou descer os valores das perdas para níveis aceitáveis (dos 12 aos 15%) nos próximos dois a três anos. Sem isso não se afigura possível alcançar nem a viabilização da actividade, nem a desejada e possível redução das tarifas, para que muitos e bons investimentos que nasçam na área da produção.

Por exemplo, a situação da Electra na cidade da Praia é crítica, pois a demanda de energia da cidade aumentou exponencialmente nos últimos anos e em simultâneo as perdas. Segundo o Relatório Anual da Electra (2007), da energia produzida em 2007 no total de 268.518.337 Kwh foi distribuída para a rede pública 236.232.450 Kwh cerca de (88%). A energia consumida internamente pela empresa rondou-se os 32.285.887 Kwh. As perdas de energia eléctrica (perdas técnicas e comerciais) passaram de 21,5 % no ano 2006 para 25,2 % no ano 2007.



Esquema 2 - Percentual das Perdas

Mais precisamente na ilha de Santiago, os valores das perdas atingiram quase os 34,6% no ano 2008. Assim as perdas de energia eléctrica, que atingiu níveis muito preocupantes,

levando já a Electra, juntamente com as entidades governamentais competentes a uma acção concertada para a resolução do problema.

Neste contexto a motivação do presente trabalho surge do interesse de investigar detalhadamente o sistema de distribuição de energia eléctrica e propor uma abordagem com o intuito de identificar perdas nos Postos de Transformação, zonas geográficas bem como consumidores fraudulentos.

1.3. Objectivos Gerais e Específicos

O **objectivo geral** do presente trabalho é desenvolver um sistema de suporte a detecção de perdas de energia eléctrica que se baseia na descoberta do conhecimento a partir da base de dados dos postos de transformação e dos clientes, identificando padrões de informação que viabilizem a descoberta de indícios, evidências ou, pelo menos, suspeitas da ocorrência de perdas, baseados na análise do comportamento dos consumidores (registos internos da base de dados) e das fontes externas.

Os **objectivos específicos são**:

- Avaliar e analisar os registos das BD para identificar aqueles que poderiam compor as dimensões de informação relevantes para o caso em estudo.
- Verificar e diagnosticar a situação dos Postos de Transformação (PT's).
- Pesquisar, adquirir e aprofundar conhecimentos sobre *Data Mining, Data Warehouse e OLAP*.
- Elaborar medidas de análise composta de índices, indicadores, parâmetros e referências, contemplando a necessidade de agregação de dados e informações.
- Construir os padrões de informação que apontem as evidências, suspeitas ou indícios de perdas.

Partindo de uma análise de dados relativos aos registos dos consumidores da base de dados dos consumos dos clientes e dos transformadores tentaremos identificar padrões na informação que permitam auxiliar a detecção de perdas comerciais com o intuito de nortear tomadas de decisões de gestão estratégicas, nomeadamente em relação às políticas de controlo e prevenção de perdas de energia eléctrica.

O modelo baseia-se na aplicação da tecnologia de extracção de conhecimento *Data Mining* na base de dados do registo e consumo dos clientes, a qual possibilita a descoberta de correlações e informações implícitas, dificilmente identificáveis utilizando as técnicas convencionais de análise do comportamento do perfil dos consumidores.

São aplicadas tecnologias como o *Data Warehouse*, *OLAP* e *Data Mining* para identificação de perdas de energia eléctrica proporcionando tomadas de decisões eficientes. A tecnologia *Data Warehouse* organizará e disponibilizará os dados, visando facilitar as tarefas de análise recorrendo a ferramentas *OLAP* e *Data Mining*.

Por sua vez a tecnologia *OLAP* permitirá fazer agregações dos dados contidos na *Data Warehouse*, gerando informações úteis e oferecendo uma análise mais detalhada para a identificar os PT's e as zonas geográficas com maior percentual de perdas. Enquanto que a tecnologia *Data Mining* visa encontrar clientes com possíveis fraudes de energia eléctrica e anomalias de leituras e de facturação do consumo.

Os resultados obtidos serão utilizados para identificar padrões e descobrir indícios de perdas comerciais e fraudes de energia eléctrica. Neste contexto, este estudo, tem como pergunta de partida**:**

Quais os padrões de informação existentes na Base de Dados da Electra, que permitem a descoberta de indícios, evidencias ou, pelo menos, suspeitas da ocorrência de perdas de energia eléctrica?

1.4. Organização do Trabalho

O foco deste trabalho está na detecção de perdas na rede eléctrica da Electra. Por conseguinte, este trabalho estrutura-se em seis capítulos. O primeiro capítulo tem a ambição de dar ao leitor uma visão sobre as **questões introdutórias**, descrição do problema, os objectivos gerais e específicos, a motivação bem como justificativa e a importância do trabalho utilizados no desenvolvimento do mesmo, proporcionando ao leitor um caminho para uma boa interpretação do mesmo.

O segundo capítulo descreve a arquitectura do sistema eléctrico**,** procedimentos para a prevenção e controlo de perdas, as formas de detecção de perdas e fraudes de energia eléctrica relacionando-as com os pontos de medição da rede eléctrica. Ainda no segundo capítulo descrevem-se as melhores práticas ou abordagens utilizadas na detecção de perdas

e as tecnologias de combate às mesmas. Neles são abordadas questões relativas aos dados relevantes para o estudo, uma visão geral do panorama actual internacional e em Cabo Verde quanto às melhores práticas e tecnologias de combate utilizadas.

No terceiro capitulo são apresentadas as principais características, potencialidades e conceitos das Tecnologias de Extracção do Conhecimento bem como uma breve descrição da Inteligência dos Negócios nas empresas.

Depois apresenta as fases do Sistema de Descoberta do Conhecimento em Base de Dados bem como as linhas gerais do processo de construção do sistema, descrevendo pormenorizadamente as tecnologias relacionadas mais importantes como *Data Warehouse, Olap* e *Data Mining*.

Será dada uma atenção especial as áreas de aplicação da tecnologia *Data Mining* exemplificando e comparando estudos de diferentes autores em diferentes áreas de aplicação.

O capítulo quatro trata especificamente da apresentação do modelo de Sistema de Informação proposto a detecção de perdas na rede eléctrica baseado na aplicação de tecnologias de extracção de conhecimento. Nele é descrito o modelo e os resultados obtidos da exploração e tratamento dos dados utilizando as tecnologias de extracção do conhecimento.

Para finalizar, o Capítulo (5) foi reservado para a apresentação das **conclusões**, e recomendações para trabalhos futuros. Em seguida, são apresentadas as referências bibliográficas ao material utilizado para a construção do modelo teórico deste trabalho. São apresentadas também, em anexo, cópias de documentos referenciados no texto.

2. Caracterização do Sistema Actual e Experiências similares noutras Instituições

Este capítulo descreve a empresa de produção e distribuição de energia eléctrica em Cabo Verde, Electra, e apresenta casos de estudo em instituições internacionais relacionados com a prevenção e combate às fraudes de energia eléctrica. Também contempla outros aspectos relevantes para o estudo como as melhores práticas e tecnologias utilizadas nessa área.

São investigados diversos estudos que tratam de variáveis socio-económicas e educacionais relacionadas com o problema e abordagens utilizadas pelas instituições nacionais e internacionais para diminuição das perdas comerciais que podem trazer benefícios para o caso da realidade cabo-verdiana.

2.1. Arquitectura do Sistema Eléctrico

O sistema produtor de energia eléctrica de Cabo Verde baseia-se essencialmente na exploração, em redes isoladas, de centrais eléctricas equipadas com grupos Diesel, utilizando como combustível o gasóleo e muito recentemente o fuel. Todos os grupos dispõem de sistemas de regulação de velocidade e de tensão de modo a manter as grandezas eléctricas da rede em valores próximos dos normais.

Segundo profissionais da Electra, as perdas nas redes rurais atingem os 33% e nas redes urbanas os 17%. As metas estabelecidas são de 8% para as redes urbanas e 15% para as redes rurais. As perdas administrativas (avenças, furtos, dívidas) continuam elevadas.

Um estudo realizado pelo Instituto Nacional de Estatística (INE, 2002), estimava-se que, no ano 2002, 60% da população tinha acesso a electricidade. Contudo, a tendência actual é no sentido de um crescimento rápido da população servida. As previsões no início do processo de privatização da Electra (Empresa de Produção e distribuição de Água e Energia Eléctrica) apontavam para uma cobertura de quase 90 % das famílias até 2008 (PEAS, 2002).

As linhas aéreas de média tensão fornecem normalmente energia eléctrica para áreas rurais, pequenas cidades, companhias industriais ou fábricas, enquanto que as linhas subterrâneas para os grandes centros urbanos.

A energia que sai da Central de Produção passa pelas subestações de distribuição da empresa, é medida, portanto, conhecida e facilmente controlada. O caminho a ser percorrido por esta energia é determinado por um conjunto de cabos de média tensão (MT), além de outros equipamentos como reguladores de tensão e de amperagem.

A energia eléctrica sai da subestação é distribuída para todos os postos de transformação anéis da cidade e estes por sua vez distribuída para os transformadores das diferentes regiões ou zonas. Essa energia depois de transformada entra nas portinholas de cada consumidor antes de ser consumida pelos clientes.

Porém, para a maioria dos consumidores, antes de a energia distribuída ser realmente consumida, especialmente os clientes residenciais e a maioria dos comerciais, é necessário realizar nova transformação com nova redução da tensão, ou seja, para baixa tensão (BT).

Esta transformação acontece em equipamentos que normalmente não possuem medição, os transformadores de BT. A energia é distribuída e consumida, mas não é possível determinar com precisão os caminhos realmente tomados por ela nos equipamentos e

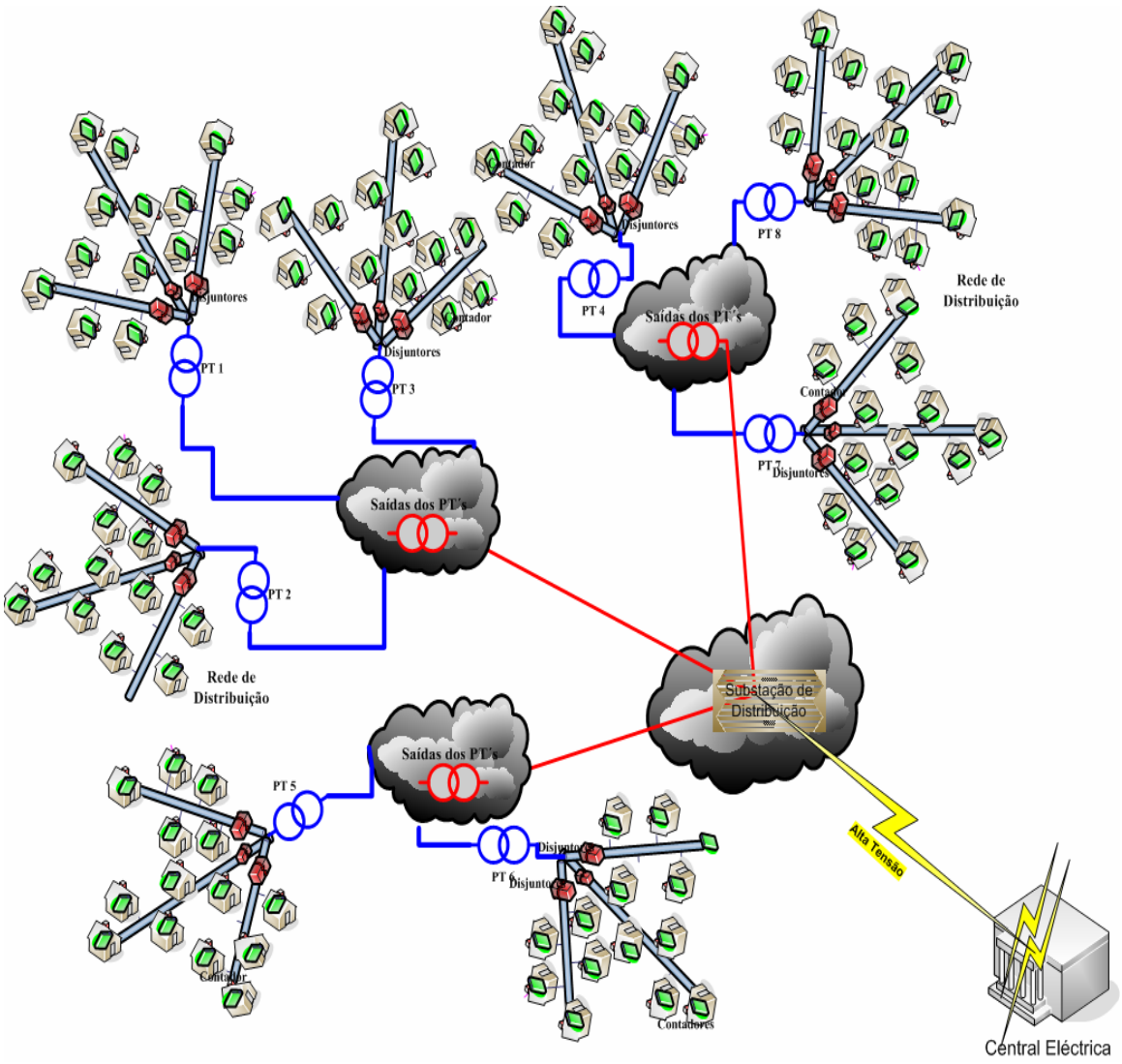
cabos. Sabe-se apenas o montante distribuído naquele transformador através dos contadores aí instalados.

A partir daí a energia é redistribuída para as portinholas para o acesso a diferentes clientes. Daí a incapacidade de se determinar com precisão as fontes de perdas comerciais em uma região atendida por um transformador. Não se sabe onde a energia passou em maior e menor quantidade, a não ser pela energia medida nas residências, casas comerciais e indústrias através dos contadores internos instalados.

A medição do consumo de energia tem por base uma estrutura de pontos de medição de geração (Central Eléctrica) e de consumo (Unidades Consumidoras), visando obter os montantes líquidos medidos de energia para cada consumidor, permitindo a contabilização e a liquidação financeira das operações. Os Pontos de medição são os locais de instalação de instrumentos (contadores de energia) para medir grandezas eléctricas afim de informar os dados gerados e de consumo. Os contadores de medição de facturação devem ser instalados nos pontos de conexão entre a subestação e o transformador e entre as portinholas e o consumidor final. Esses contadores foram desenvolvidos para facturação de energia eléctrica activa e reactiva em consumidores ligados a baixa tensão e possui 5 dígitos.

As leituras dos contadores devem ser direccionadas para as unidades comerciais, através da aquisição local das leituras em tempo integral, na qual são processadas as mesmas para contabilização e facturação. De realçar ainda que as leituras dos contadores de energia de cada consumidor são feitas mensalmente, por um leitor de uma determinada região.

O esquema 3 mostra-nos a arquitectura da rede eléctrica desde da produção de energia na Central até aos clientes finais.



Esquema 3 – Arquitectura Eléctrica

A utilização da energia distribuída é particular e imprevisível para cada consumidor. Só existe uma forma de se determinar o quanto realmente foi utilizado através da leitura do contador individual de cada consumidor.

Como não é possível realizar a leitura de todos os consumidores ao mesmo tempo, existe um ciclo de leitura, e, acompanhando esta leitura um ciclo de facturação deste consumo. Portanto, a empresa tem a exacta noção do quanto saiu das subestações e transformadores em um dado período, porém não sabe quanto foi perdido ou utilizado pelos consumidores de um transformador de uma dada região, até que este seja facturado.

A leitura nem sempre é realizada no mesmo dia, pois existe um período no qual é

possível para a empresa realizá-la. Isto provoca uma variação no montante facturado pela empresa em relação ao que foi distribuído no período, que pode também gerar uma inconsistência entre o valor medido e o valor apurado para o cálculo de perdas comercias. Outro factor que distorce a comparação entre os dois valores é uma eventual impossibilidade de leitura de alguns consumidores, provocando a facturação pela média dos últimos três meses, chamado de estimativa.

2.2. Perdas e Fraudes de Energia Eléctrica

A fraude de energia elétrica ocorre na alteração do funcionamento dos equipamentos de medição (contadores), visando redução no registo do consumo, induzindo ou mantendo a empresa em erro. As infrações ocorrem quando há troca nas ligações de medições que fazem o disco girar para trás, bloqueio do disco do contador, entre outras causas. A ligação clandestina e desvio de energia são citados como exemplos de fraudes e furtos muito comuns em Cabo Verde. Enquanto que as perdas, no âmbito da energia eléctrica, pode ser definido pela diferença entre os valores fornecidos e medidos de energia eléctrica em uma região em um dado intervalo de tempo.

O tema das perdas eléctricas tem um grande destaque nos estudos de planeamento, principalmente nos últimos anos devido aos inúmeros programas de conservação de energia, já que as mesmas representam uma grande parcela dos custos dos sistemas de distribuição de energia eléctrica.

De salientar que, as principais causas das perdas comerciais são:

- Falta de medição;
- Falhas no registo;
- Erros de medição e na facturação;
- Fraude interna;
- Iluminação pública;
- Desvio de energia;
- Ligação clandestina e fraude.

Das causas citadas as cinco primeiras estão sob o controlo da empresa, Electra, entretanto as três últimas estão fora do controlo da mesma.

Normalmente, é difícil detectar o **desvio de energia**, que é executado com bastante

cuidado e exige da Electra o desenvolvimento de treinos específicos e utilização de equipamentos especiais para facilitar a sua identificação.

Além das acções de fiscalização que devem ser utilizadas para o combate à realização deste tipo de irregularidade, pode-se utilizar também as denúncias feitas por consumidores, mesmo que de forma anónima.

As **ligações clandestinas** realizadas sem autorização da empresa ocorrem em locais onde a empresa já possui rede de distribuição atendendo os consumidores da região ou por dificuldades para a realização de uma ligação via meios normais da empresa ou devido à intenção do consumidor de obter o fornecimento sem o pagamento do consumo. Com o intuito de localizar estas ligações devem ser utilizadas as informações de leitores de contadores e equipas de manutenção da rede e de atendimento de emergência.

O aumento das fraudes de energia eléctrica na empresa de produção e distribuição de energia eléctrica – Electra, tem sido motivo de grande preocupação. Os principais motivos desse aumento, são ocasionados por ligações clandestinas, falhas na medição, erros de leituras e principalmente com o desvio ou furto de energia.

Actualmente para identificar essas situações são realizadas inspecções nas unidades consumidoras. Devido ao número elevado de unidades, tais inspecções são efectuadas sem uma análise eficiente do comportamento dos clientes.

As perdas comerciais (furto, anomalias e fraude de energia eléctrica) são encaradas como um problema mundial prejudicando a sociedade e acarretando aumento na tarifa de fornecimento e injustiça social, além de causar, em alguns casos, acidentes fatais.

As perdas comerciais causadas por ligações irregulares são as que constituem fraude e, portanto, são os alvos principais deste estudo.

2.3. Monitorização do Consumo e Detecção de Perdas

A estrutura de pontos de medição de geração e de consumo gera diariamente um volume de dados proveniente de vários contadores de medição de energia activa e reactiva desde a saída da central eléctrica até o consumidor final, os quais registam diariamente as suas demandas, como:

- Contador de Produção – Contador que regista diariamente a quantidade produzida, o factor de potencia e a tensão.
- Contador na saída da Subestação – contador que regista a quantidade de energia que sai para os transformadores.
- Contador na saída dos transformadores – contador que regista a quantidade de energia que sai para os consumidores finais de cada região.

A informação assim obtida permite ter o conhecimento da quantidade de energia produzida e distribuída.

Convêm realçar que os dados vindos da base de dados de consumo é diferente das recolhidas nos contadores de medição de geração da energia eléctrica, visto que a leitura do contador do consumidor final é feita mensalmente e a leitura dos contadores de geração diariamente. Assim a informação da base de dados da Electra está dividida em dois grupos:

- Informações comerciais, que se referem a identificação dos consumidores, local ou região de facturação, tipo de actividades.
- Informações técnicas dos alimentadores, transformadores, redes, postos e contadores.

A contagem dos clientes domésticos é ainda hoje realizada manualmente. A consequência imediata dessa leitura manual é o não conhecimento em tempo real e útil, por parte dos fornecedores e clientes, das quantidades efectivamente fornecidas, o que implica um complexo e trabalhoso processo de facturação, eventualmente com evidentes problemas de erros de dados, na resolução de anomalias e na detecção de perdas de energia.

Todos os dados são sujeitos a uma validação automática em várias fases do processo (sempre que há uma leitura de dados ou armazenamento na base de dados, sempre que os dados são editados ou são efectuados novos cálculos). Essa validação dos dados e da facturação associada entra em conta com perfis temporais de consumo, prognósticos e previsões, estando igualmente previstos métodos de substituição de valores.

As perdas técnicas em sistemas eléctricos de potência são perfeitamente equacionáveis, e, portanto, viáveis de serem administradas.

A redução de tais perdas depende, fundamentalmente, da tecnologia em utilização, da qualidade dos serviços de manutenção efectivados, da ampliação do sistema eléctrico em consonância com a evolução do mercado consumidor e do modo de operação dos sistemas.

Actualmente, o combate a perdas de energia eléctrica é feito através de inspecções em zonas ou localidades geográficas sem terem uma directriz de busca, em que o total da energia fornecida é muito diferente do total da energia medida nas residências e nos estabelecimentos comerciais da referida localidade o que torna o processo lento e dispendioso.

Essas localidades são delimitadas pelas subestações, pelos transformadores e pelas portinholas eléctricas que fazem a distribuição da energia até ao consumidor final. O total da energia fornecida à localidade é medida na saída das subestações e dos transformadores. O total da energia consumida, que efectivamente é facturada pela empresa, é verificado nos contadores instalados em cada ponto de consumo.

Assim que uma localidade é escolhida para ser inspeccionada, todas as unidades consumidoras da referida localidade são visitadas para verificação de anomalias nas ligações eléctricas.

De acordo com inspecções feitas no final do segundo semestre de 2008 de algumas zonas, a taxa de sucesso é calculada pela divisão do número de inspecções em que é identificada fraude, pelo total de inspecções realizadas nas localidades. A taxa de sucesso do processo de inspecção foi muito baixa em zonas em que os postos de transformação estavam sempre sobrecarregados provocando quedas e corte de tensões frequentes.

As inspecções no terreno nem sempre conseguem obter o flagrante de todas as perdas devido ao tempo de demora, alto custo envolvido e devidas as leis vigentes.

As perdas comerciais causadas por ligações irregulares são também fraudes que são alvo de inspecção/detecção. Estima-se que o montante de perdas provocadas intencionalmente por consumidores em Cabo Verde atinge os 25,2% em 2007 e na ilha de Santiago 34,6% segundo o Relatório Anual da Electra (2007).

Para além do prejuízo financeiro da empresa em causa (Electra) as fraudes de energia eléctrica causadas por ligações irregulares na rede podem causar acidentes graves e incêndios, por alterar as características da rede. Por isso, essas fraudes também representam um risco à segurança pública.

2.4. Melhores Práticas ou Abordagens utilizadas na Detecção de Perdas

Hoje discute-se a melhor forma de combater as perdas utilizando diferentes tipos de tecnologias e ferramentas. Em Cabo Verde já foram feitas diversas tentativas de desenvolver mecanismos para neutralizar os clientes fraudulentos, porém, inúteis.

Algumas das técnicas existentes são baseadas na instalação de dispositivos na rede de distribuição. No Brasil, mas precisamente a Escelsa (empresa de distribuição de energia eléctrica) tem desenvolvido um aparelho electrónico para detecção de desvio de energia eléctrica, que tem seu funcionamento baseado na comparação de valores de corrente. Esse aparelho possui dois módulos, um transmissor instalado nos cabos que saem do poste e um receptor que é colocado no contador do cliente (Araújo et al, 2006). A identificação do desvio é feita pela comparação dos valores de corrente medidos nos dois módulos. A comunicação é feita via PLC (*Power Line Communication*), no qual se usa a própria rede como meio de transmissão de dados (Araújo et al, 2006).

Quando a diferença entre os valores medidos ultrapassa 10%, o módulo instalado no cliente regista a informação gravando data e hora da ocorrência. Os dados armazenados podem ser transferidos para um microcomputador através de um leitor óptico portátil. A desvantagem desse equipamento é que deve ser instalado no contador do consumidor, facilitando a ocorrência de fraudes. Apesar disso, seria um bom método de detecção de fraudes se pudesse ser instalado sem que o cliente tivesse consciência de que está sendo acompanhado (Araújo et al, 2006).

Outras soluções tentam, ao invés de detectar a perdas inibi-la. Por exemplo a Ampla (empresa brasileira), desenvolveu um equipamento que impede a ligação clandestina de consumidores. Esse equipamento é instalado junto ao transformador e age como um gerador de ruídos, que distorce a tensão e "contamina" a energia a ser consumida. A tensão modificada impossibilita o consumo para quem furta. Junto aos contadores dos clientes é instalado um filtro que elimina a distorção e deixa a energia adequada para o uso. Sem o filtro, frigoríficos e outros electrodomésticos não funcionam, e a insistência no uso pode danificá-los. Apesar de bastante eficiente no combate às ligações clandestinas, esse equipamento não impede completamente o furto de energia. Os indivíduos fraudulentos

poderiam fazer o desvio da energia após o filtro e antes do contador. Assim, torna-se necessário investir em outras tecnologias para combater o furto de energia eléctrica (Araújo et tal, 2006).

Também existe o recurso a técnicas de detecção, tipicamente através do processamento computacional dos dados dos consumidores armazenados nas bases de dados das instituições, por vezes com auxílio de inspecções ao local de consumo. Para viabilizar o procedimento de inspecção, é necessário identificar previamente os consumidores que apresentam comportamentos suspeitos. Dessa forma, a inspecção é realizada de modo mais direccionado e não indistintamente. Por isso, muitos estudos se destinam ao objectivo de identificar consumidores suspeitos, a fim de investigá-los posteriormente.

O trabalho realizado por (Eller, 2003) apresenta uma arquitectura capaz de gerir perdas comerciais em energia, propondo o uso de Redes Neuronais para a identificação de possíveis clientes fraudulentos através das técnicas de classificação e segmentação. Os resultados apresentados mostram uma melhoria na identificação de clientes fraudulentos em relação a outros tipos de processos.

A sugestão de (Cabral, 2004) consiste em utilizar *Rough Sets* na selecção de características relevantes para utilização em técnicas de classificação. Os resultados foram satisfatórios, mostrando que a técnica utilizada revela-se aplicável na área de detecção de fraudes.

O trabalho apresentado em (Queiroga, 2005) analisa o uso de *Data Mining* na determinação de padrões que indiquem a possibilidade de fraudes em energia eléctrica. Os resultados obtidos com o uso de técnicas de classificação mostraram uma melhoria significativa na identificação de fraudes em relação aos métodos tradicionais.

Em (Reis, et al, 2004) é apresentado um sistema de pré-selecção de clientes de energia Eléctrica para inspecção, com o objectivo de detectar fraude e erros de medição. A partir da base de dados de uma empresa de distribuição de energia eléctrica, foram seleccionados 5 atributos (entre os 52 disponíveis) e 40.000 registos (de um total de 600.000). O sistema é baseado em uma árvore de decisão Cart, a qual foi treinada com 20.000 registos seleccionados aleatoriamente. O teste do sistema com os 20.000 registos remanescentes resultou em uma taxa de acerto de 40% para clientes fraudulentos e 35% a mais que a taxa alcançada pela empresa em questão.

Combinações de soluções que implicam instalações de dispositivos combinados com

sistemas de informação também existem, merecendo destaque ente outras, a utilização de medição externa e de contadores pré-pagos, a blindagem de cabos e o desenvolvimento de novos tipos de contadores de medição de energia e de software que emprega a inteligência artificial para obter informações úteis para aumentar a eficácia das inspecções.

Analisando as tecnologias actualmente utilizadas pelas empresas internacionais para o combate a fraude pode-se dizer que existem vários os indicadores para o auxilio no processo de detecção de fraudes, das quais destacamos:

- Identificação das áreas criticas – identificação clara do local com maior incidência de perdas comercias.
- Balanço energético – Diferença entre a energia medida pelos contadores instalados junto aos postos de transformação e a energia medida pelos contadores instalados nas unidades consumidoras conectados aos referidos transformadores.
- Sistema de facturação – extremamente importante acção no combate às perdas comerciais, pois, permite a inserção nos seus sistemas de facturação de ferramentas que possibilitam a obtenção e a gestão de informações precisas referentes a variações acentuadas no consumo de energia de unidades consumidoras que, por exemplo, pode ser realizada através de *Data Mining.*

No contexto geral podemos frisar que algumas, empresas tem desenvolvido equipamentos com a função de tornar a violação do contador de energia eléctrica mais difícil e mais robustos numa tentativa de reduzir as fraudes. Outros têm desenvolvido equipamentos que são instalados junto aos contadores e transformadores de energia que auxiliam na detecção e inibição do furto de energia eléctrica. Podemos denotar ainda que nos últimos anos muitas empresas têm investido em pesquisas e desenvolvimento de métodos de detecção de perdas através do processamento computacional dos dados dos consumidores armazenados em suas bases de dados.

3. Tecnologia de Extracção do Conhecimento à Detecção de Suspeitas de Perdas

Neste capítulo são apresentadas as principais características potencialidades e conceitos do Sistema Baseado em Conhecimento bem como uma breve descrição do Sistema de Suporte a Decisão nas empresas, mostrando sua relevância para o actual mercado competitivo e tecnológico.

Em seguida, faz-se uma abordagem aos Sistemas de Descoberta do Conhecimento em Base de Dados bem como as linhas gerais do processo de construção do sistema, descrevendo pormenorizadamente as fases mais importantes do sistema como *Data Warehouse, OLAP* e *Data Mining*.

3.1. Sistema Baseado em Conhecimento

Sistema Baseado em Conhecimento (SBC) pode ser definido como sendo um sistema computacional que representa e utiliza conhecimento para resolução de problemas (Rezende et al, 2003).

Segundo o mesmo autor, a manipulação da informação hoje em dia é um factor nuclear para as empresas que queiram diferenciar-se no mercado competitivo e perspectivar um futuro diferente. Simples bases de dados não atendem mais à demanda das empresas.

Por esse motivo a extracção do conhecimento de sistemas de informação ou bases de dados vem ganhando grande importância e interesse, sendo testemunhada pelo aumento da investigação na área e o aparecimento de tecnologias que suportam o processo de extracção e recuperação de informação (Rezende et al, 2003).

Neste contexto inserem-se os sistemas inteligentes. Estes podem variar desde simples sistemas de controlo até sofisticados *Data Warehouses* (Inmon, 2005), *Data Mart* (Inmon, 2005),), *Bussines Intelligence* (BI) (Inmon, 2005), e ferramentas *OLAP (Online Analytical Processing)* (Inmon, 2005), A principal característica comum é que estes sistemas procuram fornecer visualizações mais precisas e mais rápidas que evidenciem a informação mais relevante. Naturalmente a sua utilização e configuração depende das necessidades de quem os utilizar (especialistas em SBC) de forma a integrar o conhecimento da área do negócio.

Goldschmidt (Goldschmidt et al, 2005) quando se refere a estas tecnologias de extracção de conhecimento, subentende ferramentas computacionais capazes de realizar consultas e análises complexas em grandes volumes de dados, a partir da utilização de processos de extracção de informações implícitas. Estes dados, relações, informações genéricas, relevantes e previamente desconhecidas podem ser extraídos a partir da formulação prévia de hipóteses ou não.

Consequentemente, nota-se que há vantagens na elaboração e utilização de um SBC nomeadamente na representação do conhecimento em um sistema computacional (gestão das regras) e na análise do custo beneficio na sua aplicação. Todavia também existem dificuldades e limitações, a começar pela sua base de conhecimento ou seja, um SBC fica restrito ao conhecimento existente em sua base e ao conhecimento que vier a ser adicionado.

Dentro desse contexto ainda podemos descrever que um dos maiores desafios na construção de um SBC é a aquisição do conhecimento, a sua manutenção bem como a dificuldade de se avaliar ou prever como será o desempenho do sistema para tratar os casos reais.

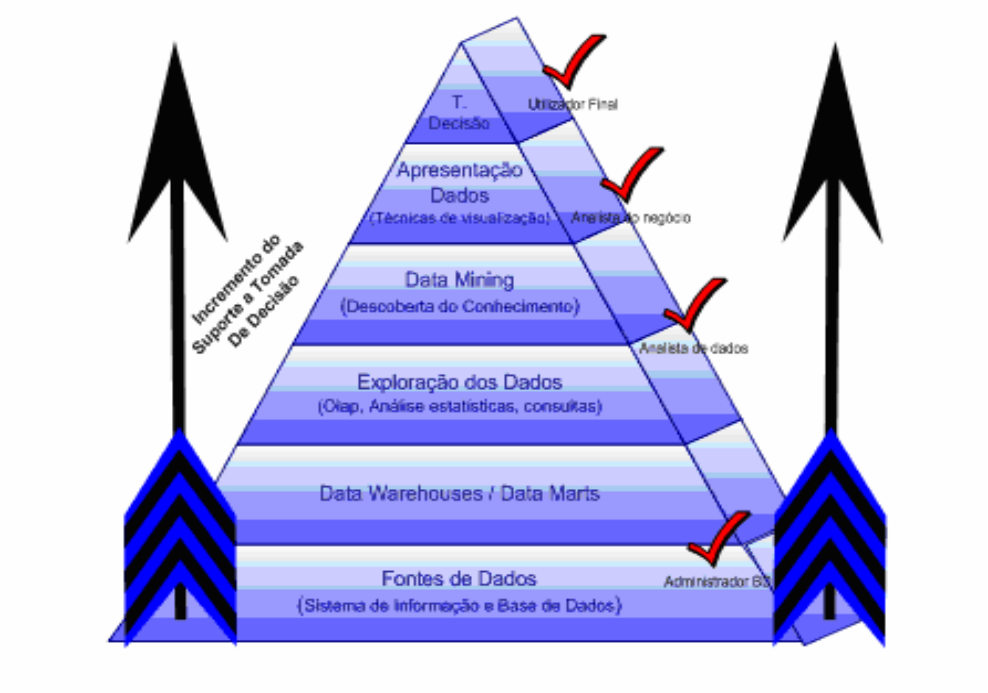
3.2. Inteligência de Negócios

Na perspectiva de (Rezende et al, 2003), actualmente as empresas de qualquer ramo de actividade enfrentam um mercado que a cada ano se torna mais competitivo, exigindo soluções mais eficientes e eficazes suportadas em sistemas de informação (base de dados, equipamentos, métodos, etc) que possam oferecer maiores vantagens e auxiliar assim as empresas de grande porte a enfrentar os desafios que surgem constantemente.

Neste contexto surge o conceito de Inteligência dos Negócios. Na perspectiva de Rezende (Rezende, et al, 2003), a Inteligência de Negócio é um conjunto de métodos ou informações que auxiliam a definição de estratégias que aumentam a competitividade no mercado e nos negócios da empresa. Assim a Inteligência de Negócio conjuga tecnologias e métodos na busca de soluções e estratégias para:

- Melhor entendimento dos segmentos de actuação da empresa no mercado;
- Poder de identificar oportunidades;
- Promover uma melhor competência na essência da empresa;
- Responder as mudanças bruscas do mercado de uma maneira mais adequada e eficiente;
- Diminuir significativamente os custos operacionais das empresas.

A Inteligência de Negócio procura assim definir regras e técnicas para a estruturação mais adequada de grandes volumes de dados, tendo como finalidade transformar a informação em uma base de dados, independente das informações que originaram.



Esquema 4 - Inteligência de Negócios e Data Mining (adaptado de Cabena, 1998)

3.3. Os Sistemas de Descoberta do Conhecimento em Bases de Dados - *Knowledge Discovery in Databases* – (KDD)

Os Sistemas de Descoberta de conhecimento em bases de dados (KDD do inglês "*Knowledge Discovery in Databases*") podem ser vistos como processos de descoberta de novas correlações, padrões e tendências significativas por meio da análise minuciosa de grandes conjuntos de dados. Estes processos baseiam-se em tecnologias de reconhecimento utilizando padrões e técnicas estatísticas e matemáticas (Fayyad, et al, 1996).

O *Data Mining* é uma das técnicas utilizadas para a realização de KDD. Nomeadamente para a investigação e criação de conhecimento, processos, algoritmos e mecanismos de recuperação de conhecimento (Fayyad, et tal, 1996).

A aplicação do *Data Mining* torna possível a transformação de dados em informação e posteriormente em conhecimento que pode ser relevante em processos de tomada de decisão. Naturalmente, é necessário investigar qual a relevância do conhecimento num

processo de tomada de decisão, bem como eventuais impactos que este conhecimento tem nas medidas ou soluções concretizadas (Silva, 2004).

Conhecer o perfil do cliente traz uma série de benefícios para a empresa, sendo o principal deles, a capacidade de melhorar a qualidade dos seus serviços prestados. Conhecendo o público alvo é possível dispor de uma melhor estratégia de marketing e com isto obter resultados mais significativos com a venda de serviços.

3.3.1. O Processo de KDD

O desenvolvimento de um KDD é uma tarefa muito complexa, principalmente por, há partida o resultado e benefícios deste serem uma incógnita. Portanto é aconselhável o uso de uma metodologia completa e sistemática que permita aumentar as hipóteses de sucesso (i.e. beneficio provinda da utilização de KDD). Vários autores apresentam várias metodologias e processos de descoberta do conhecimento em etapas. Neste trabalho vamos apresentar a metodologia defendida por (Silva, 2004).

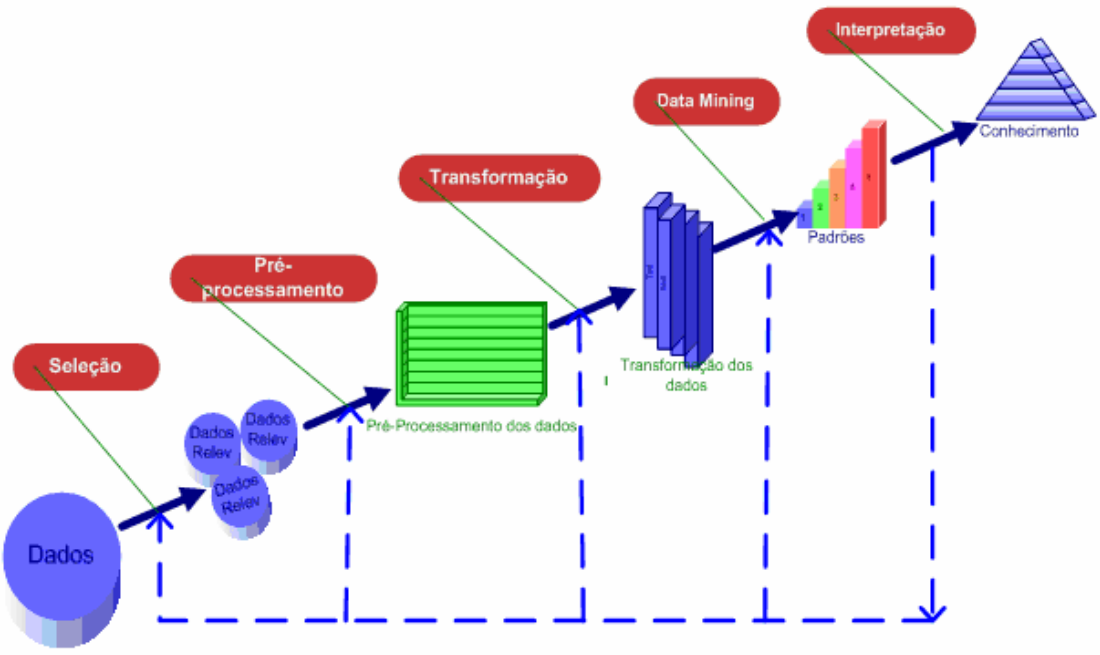
O processo de descoberta do conhecimento (KDD) é um conjunto de actividades contínuas que compartilham o conhecimento descoberto, a partir de bases de dados muito grandes. Esse conjunto é composto de etapas:

- Selecção dos dados;
- Pré-processamento e limpeza (*cleaning*);
- Transformação;
- *Data Mining* (mineração dos dados);
- Interpretação (somente após a interpretação das informações obtidas encontrar-se-á o conhecimento necessário).

Para suportar este processo de KDD é necessário o recurso a ferramentas e técnicas de análise de dados.

O processo de preparação dos dados, no qual englobamos as etapas de selecção, pré-processamento e transformação dos dados, é descrito e agrupado de maneiras diferentes por vários autores. Diferentemente de (Fayyad, 1996), que separa as etapas de limpeza e pré-processamento da etapa de redução e transformação dos dados, para (Han, 2001) o pré-processamento engloba a limpeza, a integração, a transformação e a redução dos dados,

adaptando a proposta de (Fayyad, 1996).



Esquema 5 - Vista geral das etapas que compõe o processo KDD (Adaptado de Silva, 2004)

3.3.2. Selecção dos Dados

Antes de qualquer processo de KDD, o domínio e os objectivos do problema devem ser bem entendidos para que seja possível a selecção dos dados relevantes, para que depois possa ser seleccionado e recolhido o conjunto de dados ou variáveis necessárias.

Portanto, a primeira etapa da descoberta do conhecimento, segundo (Silva, 2004), requer o conhecimento do domínio do problema e a selecção dos dados que servirão de base para a descoberta do conhecimento. Este processo iterativo implica revisões regulares e é extremamente importante para o início dos trabalhos.

O conhecimento do domínio do problema é fundamental porque é escolhido o conjunto de dados pertencente a este domínio contendo todas as possíveis variáveis e registos, e o envolvimento de especialistas no domínio é fortemente recomendado. Pode-se dizer que o processo de selecção é bastante complexo porque os dados podem vir de diferentes fontes como *Data Warehouse*, Excel e possuir diversos formatos.

3.3.3. Limpeza ou Pré-Processamento dos Dados

Esta pode ser considerada a parte crucial no processo, pois a qualidade dos dados vai determinar a eficiência dos algoritmos de *Data Mining*. Nesta etapa deverão ser realizadas tarefas que eliminam dados redundantes e inconsistentes, recuperação de dados incompletos como por exemplo, os dados seleccionados na fase anterior contêm muita informação que por vezes é repleta de ruído que necessita de ser limpo. Por eliminação de ruído entenda-se de um modo geral, a eliminação de registos duplicados **(erros e valores estranhos)**, a solução de problemas de registos com dados incompletos e campos com dados errados **(valores de atributos ausentes)**, a correcção de duplicação de conteúdo através de erros de digitação.

Quando os dados têm origens distintas, modelos distintos, é necessária uma integração dos dados e modelos com o objectivo de obter dados mais fiáveis segundo um único modelo idealmente coerente.

Na perspectiva de (Engels, 1998, citado por Silva, 2004), a qualidade da preparação dos dados pode aproximar ou distanciar os resultados do processo de *Data Mining* da solução ideal.

3.3.4. Transformação dos Dados

Após serem seleccionados, limpos e pré-processados os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos de aprendizagem possam ser aplicados. Estes dados precisam passar por um processo de redução, pois geralmente, nesta fase, a quantidade de informação disponível ainda é muito grande para ser trabalhada com eficiência. Este processo de redução pode ser suportado em mecanismos de representação eficiente dos dados e através de critérios para a redução da quantidade de atributos (seleccionando apenas os realmente necessários), do conjunto de dados usado para treino por amostras (*sampling*) entre outras técnicas. No final do processo de redução da base, pode ainda ser necessárias transformações sobre os dados (Fayyad, 1996).

Como sugere Han (Han, 2001) estas transformações podem envolver limpeza (por

remoção de discrepâncias), generalização (através da substituição de dados básicos por conceitos mais elaborados), normalização, transformações específicas ou construção de atributos (através da construção de novos valores derivados dos valores básicos para o auxílio da mineração).

De salientar que nesta fase é comum o uso de *Data Warehouses* (DW) e *Data Mart* (DM), já que estas facilitam a organização dos dados de forma eficiente. Segundo (Inmon, 2005), DW é um depósito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar as decisões de gestão enquanto que DM é uma DW de menor capacidade e complexidade utilizado para atender uma unidade específica de negócios.

3.3.5. Data Mining

Existem um conjunto de técnicas utilizado para descobrir o conhecimento em bases de dados robustas (KDD). Existem também muitas definições sobre *Data Mining* de diferentes autores, mas talvez a definição que destaca por ser mais completa é o de Usama Fayyad (Fayyad et al, 1996):

"…o processo não trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis"

O processo é não trivial já que alguma técnica de inferência é envolvida, ou seja, não é apenas um processo de análise computacional directa sobre os dados. Os padrões descobertos devem ser válidos com algum grau de certeza, novos para o sistema e para o utilizador, potencialmente úteis que traga algum benefício e compreensíveis após a interpretação.

Das etapas enunciadas anteriormente o *Data Mining* é uma das principais. No *Data Mining* é essencial a exploração e o inter-relacionamento dos dados. Este aspecto diferencia o *Data Mining* de outros processos de KDD defendidas por outros autores, suportados em ferramentas de análise disponíveis que utilizam métodos baseados na verificação, isto é, o utilizador constrói hipóteses sobre inter-relações específicas e, então, verifica ou contesta, através do sistema (Fayyad, 1996).

Estes processos dependem fortemente da intuição e habilidade do analista em propor hipóteses interessantes, em manipular a complexidade do espaço de atributos e em refinar a análise, baseada nos resultados de consultas as bases de dados potencialmente complexas.

Para suportar todo o processo de extracção de conhecimento dos dados o *Data Mining* apoia-se em modelação matemática para auxiliar a identificação de padrões nos dados observados. Naturalmente se os resultados do *Data Mining* representam ou não conhecimento num dado domínio depende do analista ou especialista que interage com os processos KDD (Berry e Linoff, 1997).

Este último aspecto é bastante relevante pois evidencia o facto do KDD ser dependente de factores humanos (i.e. através dos especialistas humanos) uma vez que estes condicionam a validação dos seus resultados (Berry e Linoff, 1997).

Na secção *Data Mining* - uma visão detalhada, iremos ver com mais detalhes o processo de *Data Mining.*

3.3.6. Interpretação e Avaliação dos Resultados

A interpretação dos resultados surge como o término da etapa anterior e consiste em analisar o resultado obtido, para identificar se é satisfatório ou se há necessidade de retornar a etapas anteriores para reformulá-las.

Nesse caso, os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas. Porém, estas formas (interpretação de padrões) devem possibilitar uma análise criteriosa para identificar a necessidade de retornar a qualquer uma das etapas anteriores do processo de descoberta do conhecimento.

Esta apresentação das actividades pode sugerir que exista uma trajectória linear do processo de descoberta do conhecimento. No entanto, isso geralmente não se verifica, uma vez que em cada etapa pode ser identificada a necessidade de retorno para cada uma das etapas anteriores. Por exemplo, se na actividade de codificação, ou mesmo na de *Data Mining*, é identificado que os dados não estejam plenamente consistentes, ou se for verificada a necessidade de um dado que não havia sido previsto anteriormente, isso pode levar ao retorno à fase de consistência ou mesmo de selecção dos dados.

3.4. Data Warehouse

Data Warehouse é uma base de dados de grande porte, resultado da junção de vários sistemas de base de dados ou técnicas que aplicadas em conjunto, servirão para gerar um sistema de dados, que tem por objectivo fornecer suporte na criação de relatórios, visando gerar informações que auxiliam nas tomadas de decisões de uma empresa (Corey, 2001).

Um *Data Warehouse* tem por objectivo oferecer organização, gerir e integrar bases de dados, assim como ferramentas de exploração dos mesmos, para se obter vantagens competitivas no mercado. É construído tendo como base, outras bases de dados operacionais que podem estar implementados em diferentes plataformas na mesma organização e é utilizado, geralmente, em aplicações de suporte à tomada de decisão.

Segundo (Gardner, 1998), a realização de um *Data Warehouse* é considerada um dos primeiros passos para tornar viável a análise de grande quantidade de dados no apoio ao processo decisório.

O objectivo básico é criar um repositório, que contenha dados limpos, agregados e consolidados podendo ser analisados por varias ferramentas (*OLAP e Data Mining*). Tais ferramentas apresentam facilidades para a realização de consultas complexas em bases de dados multidimensionais. Este processo objectiva integrar e gerir dados extraídos de diversas fontes, com o propósito de ganhar uma visão detalhada de parte ou de todo um negócio (Gupta, 1997).

As ferramentas utilizadas para analisar um *Data Warehouse*, normalmente são orientadas a consultas, ou seja, são dirigidas pelos utilizadores e especialistas, os quais possuem hipóteses que gostariam de comprovar. Esta abordagem que depende do utilizador ou especialista para formular as perguntas a base de dados poderá inibir que padrões escondidos nos dados sejam encontrados de forma inteligente, uma vez que o utilizador não terá condições de imaginar todas as possíveis relações e associações existentes em um grande volume de dados.

Existem três tipos principais de processamento usados com o *Data Warehouse* (Han, et, tal, 2006):

- Processamento de informação (suporta consultas, análise estatísticas e relatórios),
- Processamento analítico (ferramentas *Olap* e suas operações);

- Processamento de *Data Mining* (descoberta de conhecimento automático);

A concepção de um *Data Warehouse* destina-se geralmente a fornecer uma única origem aos dados para todas as actividades de apoio a decisão. O propósito de construir uma espécie de *Data Warehouse* departamental pequeno e de uso especial, adaptado à finalidade imediata, é uma solução aos problemas encontrados com os *Data Warehouses* de grande porte, visto que desta forma é possível o acesso mais rápido aos dados, ao contrário do que aconteceria se eles tivessem que ser sincronizados com todos outros dados a serem carregados no *Data Warehouse* completo. Essas considerações levaram ao conceito de *Data Marts*.

Data Mart é um *Data Warehouse* departamental, ou seja, um *Data Warehouse* construído para uma área específica da organização utilizado para atender uma unidade específica de negócios (Inmon, 2005). *Data Mart* facilita a tomada de decisões a um nível departamental e permite organizar dados segundo modelos de dados relacionais ou multidimensionais não voláteis (Kimball, 1998).

São uma alternativa simples e menos dispendiosa, pois na verdade são pequenos *Data Warehouses* colocados por sectores. Os custos e o tempo de desenvolvimento são inferiores em comparação a um projecto de *Data Warehouse* e gradualmente a empresa pode implementar outros *Data Marts*, até resultar num *Data Warehouse*.

Ainda sobre o pensamento de Inmom (Inmon, 2005), a escolha de muitos *Data Marts* e um único *Data Warehouse* é sinal de algumas controvérsias entre os pesquisadores e especialistas. Uma boa parte dos especialistas defende a implementação de *Data Mart* como fase inicial e existe uma unanimidade de especialistas alertando o utilizador que em momento algum ele pode esquecer o modelo corporativo, sob o risco de obter sérios prejuízos no futuro.

3.5. Análise dos Dados com a Ferramenta OLAP

OLAP (*On-Line Analytical Processing*) é uma ferramenta de acesso ao *Data Warehouse* que representa um conjunto de processos projectados para suportar a realização de operações de análise e consultas de dados. Os sistemas *OLAP* ajudam analistas e

executivos a sintetizarem informações sobre a empresa, através de comparações, visões personalizadas, análise histórica e projecção de dados em vários cenários (Inmon, 2005).

A característica principal dos sistemas *OLAP* é permitir uma visão de múltiplas dimensões dos dados de uma empresa. O utilizador visualiza hierarquias e navega pelas dimensões definindo fórmulas associadas a membros de dimensões, tem facilidade para fazer análises, definindo agregações e cruzamentos, permitindo visualizar os dados através de múltiplos níveis de hierarquias e diferentes perspectivas (Inmon, 2005).

Podemos afirmar que uma arquitectura *OLAP* possui três componentes principais, segundo (Aurélio, et., al, 2000):

- Um modelo de negócios para análises interactivas, implementado numa linguagem gráfica que permite diversas visões e níveis de detalhes dos dados.
- Um motor *OLAP* para processar consultas multidimensionais contra o dado alvo.
- E um mecanismo para armazenar os dados a serem analisados.

De acordo com Aurélio (Aurélio, et., al, 2000), a tecnologia de base de dados utilizada define se o pacote é um *ROLAP*, que liga com uma base de dados relacional, ou um *MOLAP*, que se liga a um servidor *OLAP*, através de uma base de dados multidimensional e dedicado.

A diferença básica entre ferramentas *OLAP* e *Data Mining* está na maneira como a exploração dos dados é abordada. Com ferramentas *OLAP* a exploração é feita na base da verificação, isto é, o analista conhece a questão, elabora uma hipótese e utiliza a ferramenta para confirmá-la. Com *Data Mining,* a questão é total ou parcialmente desconhecida e a ferramenta é utilizada para a busca de conhecimento (Inmon, 2005).

Defendendo a teoria de (Han, 2006), a integração de *OLAP* e *Data Mining* em um ambiente KDD pode ser muito importante na medida em que a *Data Warehouse* fornece dados de alta qualidade (limpos, integrados e consistentes) necessários para a aplicação de ferramentas de análise e consultas *OLAP*, enquanto que *Data Mining* automatiza o processo de descoberta de conhecimento de padrões interessantes, análise e exploração de um grande volume de dados disponíveis nos sistemas.

Ainda de acordo com Han (Han, 2006), existem vários desafios nesta integração, porque as ferramentas de *Data Mining* devem ser refeitas pensando-se em lidar com a representação de dados *OLAP*. Esta integração é bastante promissora devido às vantagens da representação, organização e consultas multidimensionais de dados *OLAP* (visão

histórica e multidimensional dos dados, interactividade, alto desempenho, uso de operações especificas para navegação nos dados) e da análise inteligente de dados proporcionado pela tecnologia de *Data Mining*.

Muitas vezes o processamento analítico é necessário em diversas situações nas quais se deseja obter informações referentes à evolução histórica dos dados. As tecnologias *OLAP* permitem esses tipos de consultas e melhoram o desempenho de tempo em relação àquelas feitas em base de dados relacionais.

3.6. Data Mining: Uma Visão Detalhada

Com o aumento da eficiência das empresas na recolha, organização e armazenamento de grandes quantidades de dados (i.e., obtido a partir das várias operações diárias) a quantidade de informação disponível aumenta, porém isto não significa obrigatoriamente o aumento do conhecimento associado a essa informação.

Segundo (Fayyad et al, 1996), este conhecimento precioso está na verdade escondido sob um grande volume de dados, e não pode ser descoberto utilizando sistemas de gestão de bases de dados convencionais. A solução existe, e chama-se *Data Mining*.

Na sua perspectiva, *Data Mining* é uma das principais fases do processo de descoberta de conhecimento nas bases de dados, pois aqui se define o algoritmo utilizado para fazer a identificação dos padrões nos dados. Esta tecnologia pode ser utilizada para descrever características do passado, assim como fazer estimativas ou identificar tendências para o futuro (Sferra e Corrêa, 2003).

O Processo de *Data Mining* permite descobrir padrões e relacionamentos através da construção de modelos, que são representações abstractas da realidade. Não se deve confundir o modelo com a realidade, um bom modelo é sempre um guia muito útil para entender o negócio da organização em causa e sugerir acções que o melhorem.

De acordo com (Berry e Linoff, 1997), existem dois modelos da tecnologia de *Data Mining,* o **modelo probabilístico** que usa dados e resultados conhecidos, para desenvolver modelos que possam prever resultados a partir de diferentes dados e o **modelo descritivo** que tem a finalidade de descobrir padrões existentes nos dados e utilizá-los para auxiliar na tomada de decisões.

Evidentemente, toda a empresa que conhece o seu negócio e os seus clientes, está sempre bem informada sobre os padrões mais significativos que foram descobertos ao

longo do tempo. Segundo (Brachman e Anand, 1996), o que a tecnologia *Data Mining* pode fazer, não é apenas confirmar estas observações empíricas, mas também descobrir novos padrões, alguns até muito difíceis de serem observados empiricamente.

De realçar que estes novos conhecimentos podem trazer grandes retornos para a empresa, pois propiciam um melhoramento contínuo. Assim, obtém-se uma pequena vantagem a cada mês, a cada projecto, a cada cliente, vantagem esta que contabilizada num período maior de tempo, faz o diferencial competitivo em relação as empresas que não utilizam bem o *Data Mining.*

Data Mining requer o conhecimento das ferramentas utilizadas e dos algoritmos nos quais elas se baseiam, pois eles são directamente relacionados com a precisão e velocidade obtidas pelo modelo.

Requer também um bom entendimento dos dados, pois a qualidade dos resultados obtidos pelos algoritmos é sensível aos *outliers* (valores que são significativamente diferentes dos demais pertencentes à fonte de dados), aos atributos irrelevantes ou atributos correlacionados, (como idade e data de nascimento, por exemplo), à maneira como os dados foram codificados, etc.

Neste contexto podemos afirmar então que a principal característica da *Data Mining* é a aplicação dos algoritmos aos dados pré-processados, com o objectivo de auxiliar as empresas a gerar indicadores numéricos, gráficos e relatórios onde o analista define o que deseja obter no momento da consulta, através de aplicações que possam servir de apoio a tomada de decisão nos diferentes níveis, sejam elas estratégicos, tácticos ou operacionais.

3.6.1. Tarefas de Data Mining

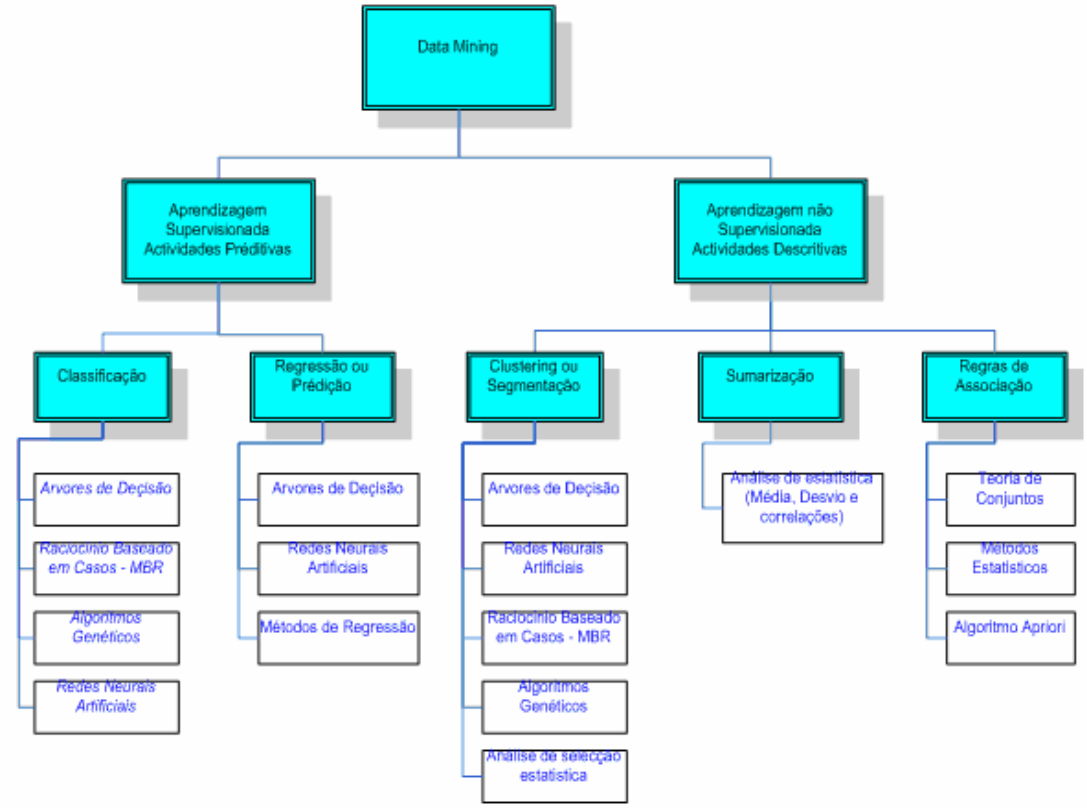
Segundo (Baptista, 2003), para explicar bem as tarefas de *Data Mining* deveremos realçar primeiro que a indução é um meio de inferência lógica que permite que conclusões gerais sejam obtidas de exemplos particulares. É caracterizada como o raciocínio que parte do específico para o geral. Ainda este autor defende que um argumento indutivo e correcto pode, perfeitamente, admitir uma conclusão falsa, ainda que as suas premissas sejam verdadeiras. Se as premissas de um argumento indutivo são verdadeiras, pode-se considerar que a conclusão é provavelmente verdadeira. Desta forma, esse recurso deve ser utilizado com muito cuidado, dado que se o número de observações for insuficiente ou se

os dados forem mal escolhidos, as hipóteses induzidas podem produzir conclusões falsas. Batista (Baptista, 2003) defende ainda que apesar disso, a inferência indutiva é um dos principais meios de criar novos conhecimentos e prever eventos futuros.

As tarefas de *Data Mining* podem ser divididas em dois tipos de aprendizagens indutivas (Resende et al, 2003):

- **Aprendizagem Supervisionada** – Consiste na generalização de exemplos ou experiências passadas com respostas conhecidas ou regras de negócios estabelecidas por especialistas usando uma linguagem capaz de estabelecer a classe de um exemplo. Pode-se dizer ainda que é onde se realizam as inferências nos dados com o intuito de realizar predições, envolvendo o uso de atributos de um conjunto de dados para prever o valor futuro do alvo (*target).* Este tipo de actividade é direccionado para a tomada de decisões.
- **Aprendizagem não supervisionada** – Consiste na identificação de comportamentos intrínsecos do conjunto de dados, sendo que estes dados não possuem uma classe especificada, ou seja, onde o tipo de actividades são descritivas e procuram padrões interpretáveis pelos humanos que descrevam os dados. Permitem a descoberta de padrões e novo conhecimento.

A escolha da tarefa é feita de acordo com os objectivos desejáveis para a solução a ser encontrada. As tarefas possíveis de um algoritmo de extracção de padrões podem ser agrupadas em actividades preditivas e descritivas, conforme indicado na figura abaixo.



Esquema 6 - Descrição de DM (Adaptado de Silva, 2004)

A Aprendizagem Supervisionada tem como objectivo identificar a classe a que pertence uma nova amostra de dados, a partir do conhecimento adquirido de um conjunto de amostras com classes previamente conhecidas. Neste tipo de aprendizagem é sempre conhecida a classe dos dados que são utilizados para treino e existe um histórico de dados que permite prever sobre dados futuros. As tarefas preditivas podem ser divididas em classificação e regressão.

Segundo (Rezende et al, 2003), a **Classificação** prediz valores discretos (classes) e permite determinar o valor de um atributo, através dos valores de um subconjunto dos demais atributos da base de dados. Ainda, tem como objectivo a construção de modelos que possa ser aplicado a dados não classificados, permitindo o agrupamento deles em classes.

Estes modelos são construídos através da análise do conjunto de treino que é retirado aleatoriamente do conjunto de dados. Após essa análise são criadas regras de classificação que são testadas com o conjunto de testes que serve para determinar a precisão das regras

de classificação.

A classificação de dados é utilizada em aplicações de diagnóstico médico, previsão de tendências, determinação de estratégias de *marketing*, etc., e vem sendo estudada em estatística, *machine learning,* redes neuronais e sistemas periciais (Weiss e Kulikowski, 1991) e é um importante tema em *Data Mining* (Fayyad et al.1996).

Por seu lado, a **Regressão ou Estimativa** tem como objectivo definir um valor numérico de alguma variável desconhecida a partir de valores de variáveis conhecidas utilizando um conjunto de dados históricos como modelo. Han, (Han et al, 2001), diz que a regressão trata principalmente de valores numéricos em detrimento das variáveis categóricas.

(Fayyad, 1996), diz ainda que a regressão é aprender uma função que permite mapear um item de dado para uma variável de predição real estimada.

Na perspectiva de (Silva, 2004), a **Aprendizagem não supervisionada** o atributo de cada classe amostra de treino não é conhecida e o número ou conjunto de classe a ser treinado pode não ser conhecido à priori. São algoritmos descritivos, pois descrevem de forma concisa os dados disponíveis, fornecendo características das propriedades gerais dos dados submetidos a *Data Mining*.

Uma vez escolhida a tarefa a ser realizada, existe uma variedade de algoritmos para executá-la. A definição do algoritmo de extracção e a posterior configuração de seus parâmetros também são realizadas nesta etapa. Por isso, a escolha de vários algoritmos para realizar a tarefa desejada pode ser feita, levando à obtenção de diversos modelos que, na etapa posterior ao processamento, são tratados para fornecer o conjunto de padrões mais adequado ao utilizador final.

As actividades descritivas trabalham com conjuntos de dados que não possuem uma classe determinada e têm o objectivo de identificar padrões de comportamento semelhantes nestes dados (Gonçalves et al, 2005). As tarefas descritivas podem ser divididas em regras de associação, *clustering* ou segmentação e sumarização.

De acordo com (Harrison, 1998), a tarefa de **associação** consiste em determinar quais factos ou objectos tendem a serem adquiridos juntos em uma mesma transacção. O exemplo clássico é determinar quais produtos costumam ser colocados juntos em um carrinho de supermercado, daí o termo 'análise de *market basket'*.

A tarefa de associação pode ser considerada uma tarefa bem definida, determinística e

relativamente simples, que não envolve predição da mesma forma que a tarefa de classificação (Freitas, 2000). Os algoritmos para as regras de associação determinam um padrão de relacionamento entre itens de dados.

A **segmentação** é um processo de partição de uma população heterogénea em vários subgrupos ou *clusters* mais homogéneos (Harrison, 1998). Na segmentação, não há classes predefinidas, os registos são agrupados de acordo com a semelhança, o que a diferencia da tarefa de classificação.

Exemplos de segmentação: agrupar os clientes por região do país, agrupar clientes com comportamento de compra similar (Goebel e Gruenwald, 1999).

Segundo Fayyad (1996), a tarefa de **sumarização** envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Um simples exemplo desta tarefa poderia ser tabular o significado e desvios padrão para todos os itens de dados. Métodos mais sofisticados envolvem a derivação de regras de sumarização.

3.6.2. Técnicas e Algoritmos de Data Mining

No âmbito do pensamento de (Harrison, 1998), não há uma técnica que resolva todos os problemas de *Data Mining*. Diferentes métodos servem para diferentes propósitos, cada método oferece suas vantagens e suas desvantagens. A familiaridade com as técnicas é necessária para facilitar a escolha de uma delas de acordo com os problemas a serem tratados. A seguir são descritas as técnicas de *Data Mining* normalmente utilizadas.

As **regras de associação** têm por objectivo descobrir importantes associações entre itens que compõem uma base de dados, de tal forma que a presença de um item numa determinada transacção, implique na presença de outro item na mesma transacção. O modelo matemático está descrito a seguir (Goebel et al, 1999):

- Uma regra de associação tem a forma geral X1 ^ ... ^ Xn => Y [C,S], onde X1,..., Xn são itens que prevêem a ocorrência de Y com um grau de confiança C e com um suporte mínimo de S e "^" denota um operador de conjunção (AND). Um exemplo desta regra pode ser que 90% dos clientes que compram leite, também compram pão; o percentual de 90% é chamado a confiança da regra. O suporte da regra leite => pão é o número de ocorrências deste conjunto de itens na mesma transacção.

A técnica de descoberta de regras de associação é apropriada à tarefa de associação. Como exemplos de algoritmos que implementam regras de associação tem-se: Apriori, AprioriTid, AprioriHybrid, AIS, SETM (Agrawal e Srikant, 1994) e DHP (Chen et al, 1996).

Uma **árvore de decisão** é um modelo preditivo que pode ser visualizado na forma de uma árvore. Cada ramo da árvore é uma questão de classificação e cada folha é uma partição do conjunto de dados com sua classificação. Dado um conjunto de dados cabe ao utilizador escolher uma das variáveis como objecto de saída. A partir daí, o algoritmo encontra o factor mais importante correlacionado com a variável de saída e identifica-o como o primeiro ramo (raiz), os demais factores são subsequentemente classificados como nós até que chegue ao último nível, a folha.

Pode-se dizer que desta forma a árvore de decisão utiliza a estratégia de dividir para conquistar, em que um problema complexo é decomposto em sub-problema mais simples e aplicando recursivamente a mesma estratégia a cada sub-problema (Goebel e Gruenwald, 1999). Uma das vantagens principais das árvores de decisão é o facto de o modelo ser facilmente explicável, uma vez que tem a forma de regras explícitas (Harrison, 1998).

As árvores de decisão são representações gráficas onde os nós representam amostras e as folhas representam categorias. Uma árvore de decisão designa uma classe numérica (ou saída) para uma entrada padrão filtrando-se a amostra através dos testes na árvore. Cada teste possui reciprocamente resultados exclusivos e exaustivos.

Quando a amostra de uma população é estudada com o objectivo de se fazer alguma inferência indutiva, as árvores de decisão são os modelos mais utilizados. Em muitos exemplos observamos árvores de decisão construídas usando apenas valores booleanos, porém não estamos limitados a implementação destas funções.

A técnica de árvore de decisão, em geral, é apropriada às seguintes tarefas: classificação e regressão. Alguns exemplos de algoritmos de árvore de decisão são: CART, CHAID, C5.0, ID-3 (Chen et al, 1996) e SPRINT (Shafer et al, 1996).

De acordo com (Harrison, 1998), **Raciocínio Baseado em Casos** também conhecido como MBR (*Memory-Based Reasoning*), tem por base o método do vizinho mais próximo. "O MBR procura os vizinhos mais próximos nos exemplos conhecidos e combina seus valores para atribuir valores de classificação ou de previsão". Tenta solucionar um dado problema fazendo uso directo de experiências e soluções passadas. A distância dos vizinhos

dá uma medida da exactidão dos resultados.

A técnica de raciocínio baseado em casos é apropriada às seguintes tarefas: classificação e segmentação. Os seguintes algoritmos implementam a técnica de raciocínio baseado em casos: *Birch*, *Clarans* (Chen et al, 1996) e Clique.

Ainda na mesma linha de pensamento (Goebel e Gruenwald, 1999), afirma que os **algoritmos genéticos** são métodos generalizados de busca e optimização que simulam os processos naturais de evolução. Um algoritmo genético é um procedimento iterativo para evoluir uma população de organismos e é utilizado em *Data Mining* para formular hipóteses sobre dependências entre variáveis, na forma de algum formalismo interno.

Os algoritmos genéticos utilizam os operadores de selecção, cruzamento e mutação para desenvolver sucessivas gerações de soluções. Com a evolução do algoritmo, somente as soluções com maior poder de previsão sobrevivem, até os organismos convergirem em uma solução ideal (Harrison, 1998).

A técnica de algoritmos genéticos é apropriada às tarefas de classificação e segmentação. Exemplos de algoritmos genéticos: Algoritmo Genético Simples, Genitor e CHC, Algoritmo de Hillis, GA-Nuggets (Freitas, 1999), GA-PVMINER (Araújo et al, 1999).

Por fim (Goebel e Gruenwald, 1999) defendem que as **redes neuronais** são uma classe especial de sistemas modelados em analogia com o funcionamento do cérebro humano e são formadas por neurónios artificiais conectados de maneira similar aos neurónios do cérebro humano.

"Como no cérebro humano, a intensidade das inter conexões dos neurónios pode alterar (ou ser alterada por algoritmo de aprendizagem) em resposta a um estímulo ou uma saída obtida que permite a rede aprender" (Goebel e Gruenwald, 1999).

Uma das principais vantagens das redes neuronais é sua variedade de aplicação, mas os seus dados de entrada são difíceis de serem formados e os modelos produzidos por elas são difíceis de entender (Harrison, 1998).

A técnica de redes neuronais é apropriada as tarefas de classificação, estimativa e segmentação. Exemplos de redes neuronais: *Perceptron*, Rede *MLP*, Redes de *Kohonen*, Rede *Hopfield*, Rede *BAM*, Redes *ART*, Rede *IAC*, Rede *LVQ*, Rede *Counterpropagation*, Rede *RBF*, Rede *PNN*, Rede *Time Delay*, *Neocognitron*, Rede *BSB*, (Braga, 2000), (Haykin, 2001).

A Tabela a seguir apresenta um resumo das técnicas de *Data Mining* aqui descritas.

Técnicas	***Descrição***	***Tarefas***	***Exemplos***
Descoberta de regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjunto de dados.	**Associação**	Apriori, AprioriTid, AprioriHybrid, AIS, SETM, (Agrawal e Srikant, 1994) e DHP (Chen e tal, 1996).
Árvores de Decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos.	**Classificação Regressão**	CART, CHAID, C5.0, Quest (Two Crows, 1999), ID-3 (Chen et al, 1996), SLIQ (Metha et al, 1996) e SPRINT (Shafer et al, 1996).
Raciocínio baseado em casos	Baseado no método de Vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança.	**Classificação Segmentação**	BIRCH (Zhang et al, 1996), CLARANS (Chen et al, 1996) e CLIQUE (Agrawal et al, 1998).
Algoritmos genéticos	Métodos gerais de busca e optimização, inspirados na teoria da evolução, onde a cada nova geração, soluções melhores têm mais chance de ter descendentes.	**Classificação Segmentação**	(Goldberg, 1989), Genitor e CHC (Whitley, 1993), Algoritmo de Hillis (Hillis, 1997), GA-Nuggets (Freitas, 1999), GA-PVMINER (Araújo et al, 1999).
Redes Neuronais Artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos destas conexões.	**Classificação Segmentação**	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation,

Tabela 1 - As técnicas de DM, adaptado de (Fayyad, 1996)

3.6.3. Ferramentas de Data Mining

Actualmente existem várias ferramentas de *Data Mining* no mercado (Han & Kamber, 2001). Muitas destas especializam-se apenas em uma única aproximação de *Data Mining*, como por exemplo em algoritmos de classificação. Outros sistemas têm uma visão mais abrangente e fornecem vários algoritmos de *Data Mining* e exploram múltiplas técnicas do processo de descoberta do conhecimento.

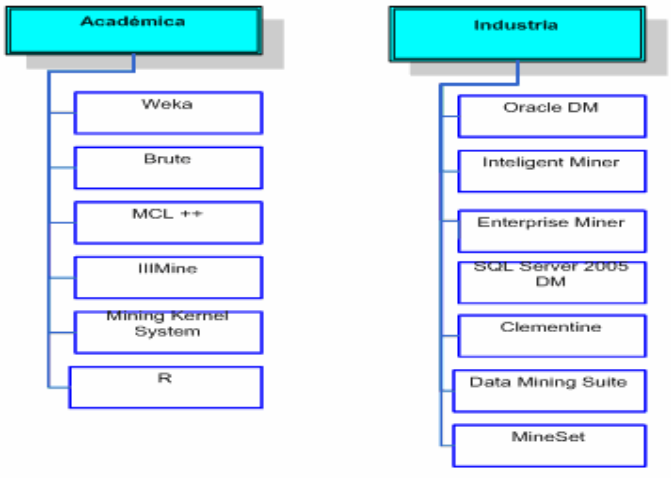
Na perspectiva de (Goebel e Gruenwald, 1999), muitas ferramentas actualmente disponíveis são ferramentas genéricas da Inteligência Artificial ou estatística. Tais ferramentas geralmente operam separadamente da fonte de dados, requerendo uma

quantidade significativa de tempo gasto com exportação e importação de dados, pré e pós-processamento para disponibilização dos resultados.

Entretanto, segundo os autores, a conexão rígida entre a ferramenta de descoberta de conhecimento e a base de dados analisada, utilizando o suporte do *SGBD* (Sistema de Gestão de Base de Dados) existente, é claramente desejável. Para Goebel e Gruenwald (1999), as características a serem consideradas na escolha de uma ferramenta de descoberta de conhecimento devem ser as seguintes:

- A capacidade de acesso a uma variedade de fontes de dados, de forma *online* e *offline*;
- A capacidade de incluir modelos de dados orientados a objectos ou modelos não padronizados (tal como multimédia, espacial ou temporal);
- A capacidade de processamento com relação ao número máximo de tabelas e atributos;
- A capacidade de processamento com relação ao tamanho da base de dados;
- Variedade de tipos de atributos que a ferramenta pode manipular;
- Tipo de linguagem de consulta.

Existem ferramentas que implementam uma ou mais técnicas de *Data Mining*. O Esquema 7 relaciona algumas dessas ferramentas, fornecendo informações tais como as técnicas implementadas de *Data Mining* e exemplos de aplicações.



Esquema 7 - Subconjunto das Ferramentas de DM, adaptado (Han et tal, 2001)

De salientar que quanto ao âmbito de aplicação, vantagens e desvantagens, visualização e formatos dos dados (*Input / Output*), estas ferramentas podem ser comparadas entre si apresentando cada uma características semelhantes ou diferentes consoante for a tarefa de *Data Mining* a ser desempenhada conforme mostra a tabela seguinte:

Ferramentas	***Âmbito de aplicação***	***Visualização***	***Formato dos Dados (Input / Output***	***Vantagens***	***Desvantagens***
Intelligente Miner (IBM)	Suporte ao OLAP e DM para todas as aplicações.	Histogramas e gráficos de linhas	ODBC / JDBC e drivers da base de dados	Quantidade de algoritmos, output gráfico e volume de dados tratáveis.	Falta de flexibilidade dos algoritmos e pouca automação
Enterprise Miner (SAS Institute)	Suporte ao OLAP e DM para todas as aplicações.	Histogramas e gráficos de linhas	ODBC / JDBC e drivers da base de dados. Não fornece Código fonte	Quantidade de algoritmos, variedade de ferramentas de análise estatística e interface visual	Difícil de Utilizar
O Darwin Data Mining Software (Oracle)	Suporte apenas DM	Histogramas e gráficos de linhas	ODBC / JDBC e drivers da base de dados	Eficiência e interface como utilizador	Falta de ferramentas na visualização da informação.
Weka (OpenSource)	Suporte apenas DM	Histogramas e gráficos de linhas	ODBC / JDBC e drivers da base de dados	Interface gráfica amigável e principais algoritmos de DM implementados	Alguns parâmetros só podem ser executados via linha de comandos (script)
MineSet (Silicon Grafics)	Suporta DM	Gráficos	ODBC / JDBC e drivers da base de dados	A robustez das ferramentas gráficas que incluem visualização em árvore, em regras e em mapa	

Tabela 2 - Caracteristicas de Ferramentas de DM, adaptado de (Pereira, 2002)

3.6.4. Selecção da Técnica de Data Mining e Ferramentas Adequadas

Na óptica de (Harrison, 1998), a escolha de uma técnica de *Data Mining* a ser aplicada não é uma tarefa fácil. Segundo ele, a escolha das técnicas de *Data Mining* dependerá da tarefa específica a ser executada e dos dados disponíveis para análise. Harrison (1998) sugere ainda que a selecção das técnicas de *Data Mining* deve ser dividida em dois passos:

- Traduzir o problema de negócio a ser resolvido em séries de tarefas de *Data Mining*.
- Compreender a natureza dos dados disponíveis em termos de conteúdo, tipos de campos de dados e estrutura das relações entre os registos.

Essa escolha pode ser baseada, também, em critérios ou esquemas para classificação das técnicas. (Chen et al, 1996), propõe os seguintes critérios de classificação, baseados em questões que devem ser levantados como por exemplo:

- Com que tipos de bases de dados trabalhar - Um sistema de descoberta de conhecimento pode ser classificado de acordo com os tipos de bases de dados sobre os quais técnicas de *Data Mining* são aplicadas, tais como, bases de dados relacionais, orientadas a objectos, dedutivas, espaciais, temporais, multimédia, ou informação baseada na Internet e bases textuais.
- Qual tipo de conhecimento a ser explorado - vários tipos de conhecimento podem ser descobertos por extracção de dados, incluindo regras de associação, regras características, regras de classificação, regras de agrupamento, evolução e análise de desvio.
- Que tipo de técnica a ser utilizada - a extracção de dados pode ser categorizada de acordo com as técnicas de *Data Mining* subordinadas. Por exemplo, extracção dirigida a dados, extracção dirigida a questões e extracção de dados interactiva. Pode ser categorizada, também, de acordo com a abordagem de *Data Mining* subordinada, tal como, extracção de dados baseada em generalização, baseada em padrões, baseada em teorias estatísticas ou matemáticas, abordagens integradas, etc.

3.7. Data Warehouse e OLAP para Data Mining

Existem benefícios reais ao utilizar *Data Mining* sobre *Data Warehouse* (Sanches, 2003). Os *Data Warehouse* organizam os dados para um efectivo processo de *Data Mining*, porém, a exploração através do *Data Mining* pode ser aplicada onde não exista nenhum *Data Warehouse*. O uso de *Data Warehouse* aumenta significativamente as hipóteses de sucesso do *Data Mining*, visto que o *Data Warehouse* dispõe de dados integrados,

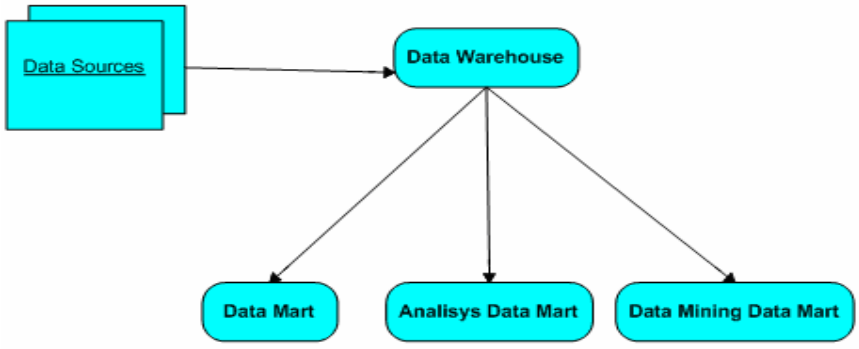
detalhados e organizados historicamente. Portanto a utilização desses tipos de dados melhora o desempenho e o resultado do processo de *Data Mining*.

Outra vantagem na sua utilização, é a similaridade entre os problemas de refinamento dos dados para *Data Warehouse* e para *Data Mining*. Daí que se os dados forem consultados directamente no *Data Warehouse*, muitos dos problemas envolvidos com a sua consolidação já terão sido resolvidos.

Os dados integrados e detalhados permitem o analista de sistema visualizar de forma rápida e fácil os dados de forma a examinar os mesmos gradualmente. Os dados históricos são importantes porque grandes quantidades de informação ficam implicitamente armazenadas. Trabalhar somente com informações actuais pode impedir que se detectam tendências e padrões de comportamento ao longo do tempo, daí que informações históricas são necessárias para o entendimento das circunstâncias dos negócios (Kimball, 2002).

De acordo com Kimball (Kimball, 2002), enquanto que o *OLAP* é dedutivo e guiado por especialistas, o *Data Mining* é indutivo e guiado pelos próprios dados. Ambas necessitam de dados limpos e consistentes. E neste caso, o *Data Warehouse* é capaz de fornecer dados para as duas tecnologias o que o torna a principal fonte de dados para a junção entre *OLAP* e *Data Mining*.

As técnicas de *Data Mining* vão ser aplicadas sobre um *Data Warehouse* ou *Data Mart* construídos a partir dos dados extraídos da base de dados existentes e outras fontes como por exemplo folhas de cálculo. Como os dados são geralmente preparados antes de serem armazenados num *Data Warehouse* também se obtêm melhorias ao nível de qualidade de informação.



Esquema 8 - Data Mining utilizando Data Warehouse

3.8. Áreas de Aplicação do Data Mining

Actualmente a área de aplicações de *Data Mining* dependem em parte do negócio no qual ele está inserido e da imaginação do especialista em *Data Mining*. Pode-se encontrar uma vasta área de aplicação do processo de *Data Mining*, segundo diferentes autores (Cratochvil, (1999), Mannila (1997) e Viveros et al, (1996)), como por exemplo, em áreas de aplicações académicas e corporativas.

3.8.1. Aplicações Académicas

No meio académico, a procura por novas abordagens de *Data Mining* está presente em muitas áreas de investigação, entre as quais (Silva, 2004):

- *Data Mining* para *Data Warehouses* – aplicações típicas são utilizadas para obter melhorias ao nível da qualidade de informação em que facilitam a organização dos dados de forma eficiente.
- *Data Mining* em base de dados espaciais - aplicável sobre sistema de informação geográfica para manipular informações geograficamente referenciadas. Pode ser aplicado sobre, satélite, dados sobre a saúde pública para o mapeamento de doenças e avaliação de riscos, ou incidir sobre estudos ambientais, vigilância territorial ou planeamento urbano, entre outros exemplos.
- *Data Mining* multimédia - extracção de padrões relevantes a partir de animações, áudio, vídeo, imagens e textos realizando por exemplo pesquisas por similaridade, análise multidimensional, preditivas, entre outros.
- *Data Mining* (séries) temporais - mercado de acções, processos de produção, experiência científica, tratamentos médicos, análises de tendências em históricos. Aplica-se também nos casos em que há um padrão persistente ou sistemático no comportamento de uma variável, que é possível captar através de uma representação paramétrica.
- *Data Mining* de texto ou *Text Mining* - muitas informações estão disponíveis em documentos (artigos de jornais ou científicos, dicionários, livros, e-mails, páginas *Web* etc). Dentre as abordagens encontramos a recuperação de informações a partir de documentos de textos sobre informação semi-estruturada

(Feldman, 1995). A tecnologia *Text Mining* permite identificar os conceitos presentes nos textos. Conceitos que permitem entender que temas estão presentes nos textos ou do que se tratam os textos.

- *Data Mining* da *Web* ou *Web Mining* - A *Web* configura-se como um repositório imenso, distribuído e global que contém uma ampla e rica colecção de informação relacionado por meio de *hiperlinks*. Seu tamanho, complexidade e dinamismo oferecem grandes desafios científicos como, por exemplo, a análise de ficheiros de Log de acesso aos servidores web (Yao et al, 2001)), Clientes Web e servidores proxy (analisando paginas armazenadas em cache). A designação *Web Mining* se refere á identificação de padrões no comportamento de uso da *web*.

3.8.2. Aplicações Corporativas

Não há na prática uma área específica para aplicação dos conceitos, tudo depende da capacidade de modelar o problema para a aplicação e da criatividade para analisar e utilizar os resultados alcançados. Atentas ao poder da informação, as empresas que recolhem grandes volumes de dados a todo o instante, investem continuamente em novas tecnologias que agreguem valor ao negócio. O efectivo emprego das técnicas de *Data Mining* atinge diferentes empreendimentos (Han, et tal, 2001):

- No processo de detecção de fraudes (fraudes de energia eléctrica, fraudes de cartão de crédito), onde se procura encontrar padrões e indicadores da existência de fraudes através da tecnologia *Data Mining*. Em instituições governamentais, onde são aplicadas na descoberta de padrões para melhorar as recolhas de taxas e impostos, detectar fraudes, bem como formular políticas públicas. Nas Instituições financeiras para detectar padrões de uso de cartão de crédito fraudulento, determinar gastos com cartão de crédito por grupos de clientes, encontrar correlações escondidas entre diferentes indicadores financeiros, etc.
- No Marketing para descobrir preferências do consumidor e padrões de compra, com o objectivo de realizar marketing directo de produto e ofertas promocionais, de acordo com o perfil do consumidor. Ajuda a descobrir grupos de clientes e

usa esse conhecimento para orientar as campanhas publicitárias e pesquisas de mercado.

- Na medicina no processo de caracterização do comportamento do paciente para prever visitas, identificar terapias médicas de sucesso para diferentes doenças, buscar por padrões de novas doenças, etc.
- Na Ciência onde técnicas de *Data Mining* podem ajudar cientistas em suas pesquisas, por exemplo, para encontrar padrões em estruturas moleculares, dados genéticos, mudanças globais de clima, podendo oferecer, rapidamente conclusões valiosas.
- Em energia e telecomunicações para previsão do consumo e de falhas em sistemas de transmissão ou de distribuição, nas vendas para a identificação de produtos, nas finanças para avaliação de riscos, previsão de flutuações nos mercados de acções, são algumas das aplicações mais comuns encontradas.

3.8.3. Aplicações de Data Mining na Detecção de Perdas

Podemos salientar que estudos e trabalhos na detecção de fraudes estão presentes em diversos ramos da actividade. Em diferentes sectores há a mesma preocupação de prevenir, detectar e combater as fraudes. No sentido geral a detecção de fraudes envolve recuperação de receitas, por esse motivo é um dos objectivos mais importantes das empresas. Quando falamos na detecção de fraudes, estamos geralmente lidando com a investigação de um número muito grande de informações.

Neste âmbito o estudo desenvolvido por Silva, (Silva et al, 2003), descreve sobre o uso de *Data Mining* em um ambiente *Web* em que foram aplicadas técnicas da descoberta do conhecimento com o intuito de investigar a relevância das informações obtidas por meio da análise dos padrões de navegação de utilizadores em *Web sites* de uma empresa provedora de acesso à Internet, descritos em arquivos *log* de um servidor *Web.* Para análise dos dados foi utilizada a ferramenta *Weka*, o algoritmo Apriori que faz uso de regras de associação. Esta análise permitiu extrair várias regras de associação, o que propiciou a identificação de padrões de comportamento de internautas ao navegarem pela Web Site da empresa. Ao ter conhecimento da frequência de acesso ao *Web Site* e quais os serviços mais procurados, a

gestão da empresa pode descobrir o perfil dos seus utilizadores e com base nisso oferecer serviços e atendimento personalizado.

Brause (Brause, 1999) demonstrou em seu trabalho para a detecção de fraudes com cartões de crédito utilizando regras de decisão sobre dados simbólicos, que não foi possível atingir o nível de confiança proposto para as experiências. No entanto no mesmo estudo quando utilizou redes neuronais sobre dados analógicos, não houve resultados, pois todos os casos eram classificados como normais. Brause, utilizou os resultados identificados pelo sistema de regras de decisão sobre dados simbólicos como entrada para redes neuronais obtendo assim resultados próximos dos desejados e reconhece ainda que recolheu dados especificamente para este fim. Nessa experiência cada uma das redes neuronais era especializada em um tipo de dados analógicos.

Burge (Burge, et tal, 1997) propõe uma ferramenta híbrida para detecção de fraudes (Brutus) na área das telecomunicações, que é composta por um módulo que utiliza redes neuronais e um outro módulo que utiliza sistemas baseados em regras. Segundo Burge, os testes iniciais utilizando dados do sector das telecomunicações, visando detectar fraudes neste domínio tiveram bons resultados apesar de serem preliminares.

Santos (Santos, 2007), apresentou um trabalho intitulado "Ambiente para extracção de informações através da mineração das bases de dados do Sistema Único de Saúde (SUS)" em que a dificuldade na extracção de informações de gestão a partir da exploração das bases de dados foi uma das questões motivadoras do trabalho. Os resultados confirmaram a coerência da informação produzida nas questões elaboradas, comprovando a capacidade do ambiente em extrair informações úteis a gestão da Saúde Publica.

Na área de saneamento Queyroi (Queyroi, 2007) desenvolveu um trabalho "Aplicação de um Modelo de Mineração de dados em um sistema de apoio a decisão para empresas de saneamento" em que utilizou árvores de decisão pelo aspecto de visualização por parte dos especialistas, proporcionando o aprofundamento das análises dos resultados. Como resultado ele, conclui que o modelo aplicado ao problema acrescentou ao conhecimento do especialista padrões de actuação que podem ser levados em conta no futuro, como por exemplo, em trocas futuras de contadores.

Segundo (Passini, 2002), as fraudes de água em Campinas contribuíram em 5% dos 26,6% de perdas na distribuição, no ano 2000, o que levou a utilização de *Data Mining* para identificar fraudes. A utilização de *Data Mining* surgiu por existirem dados históricos

armazenados há mais de dez anos, que poderiam ser investigados para descoberta de informações válidas e desconhecidas, contribuindo para a identificação de perfis de comportamento que pudessem levá-los aos potenciais consumidores fraudulentos. Para isto foram utilizados o *DB2* e o *Intelligente Miner* da *IBM* e foram elaborados três modelos de *Data Mining,* sendo dois deles baseados em segmentação neuronal e um em classificação por árvore de decisão. Para a construção dos modelos, os consumidores fraudulentos foram separados dos clientes normais. O projecto não atingiu os objectivos, os resultados não foram os esperados, mas sabia-se que o modelo ainda precisava ser melhorado.

A actividade de fornecimento de energia eléctrica é, assim como várias outras actividades, alvo comum de fraudes. Por se tratar de um serviço disponibilizado a um grande número de clientes e de difícil gestão e controle de fraudes, muitos são os casos de fraudes encontradas neste sector.

De realçar que as fraudes podem ser simples desvios instalados directamente na rede eléctrica ou manipulações mais complexas como violações ao equipamento de medição e alterações no mesmo para que este não realize correctamente a leitura.

(Eller, 2003), apresenta um estudo sobre a gestão de perdas comerciais de energia eléctrica, no qual se propõe uma arquitectura que actua na indicação de possíveis consumidores fraudulentos e ainda na gestão de perdas comerciais. Neste trabalho utilizou-se redes neuronais para identificação de possíveis consumidores fraudulentos através de classificação (Haykin, 2001), e as conclusões tiradas foram que esta é uma arquitectura com potencial, mas que ainda necessitava aprimorar os modelos para que os resultados fossem mais eficazes.

Minussi, (Minussi, 2008) elaborou um trabalho de mestrado intitulado "Metodologia de Mineração de dados de desvio de comportamento do uso de energia numa empresa de energia eléctrica" em que o objectivo principal era diminuir as perdas comerciais e maximizar os lucros. Conseguiu desenvolver o trabalho em etapas de análise e avaliação dos dados, bem como construção de um *Data Warehouse* em que foram analisadas curvas de cargas dos clientes e através dessa análise observou-se o perfil do consumo dos mesmos utilizando o algoritmo de associação apriori para fornecer padrões de indicadores de perfil dos consumidores bem como os algoritmos de árvore de decisão e classificadores *bayesianos*. Na sua perspectiva os resultados validam o método desenvolvido e implementado, permitindo sua utilização numa empresa de energia eléctrica. Os algoritmos

foram aplicados e foi traçado um comparativo dos mesmos, sendo que todos resultaram em boas respostas de *Data Mining* e poucas diferenças entre os resultados.

Um outro trabalho envolvendo *Data Mining* foi apresentado em (Queiroga, 2005), no qual um amplo estudo foi realizado na análise de perdas comercias, envolvendo diversos modelos computacionais para a geração de classificadores (redes neuronais, indutor de regras, *naive bayes*, etc).

Neste trabalho de Queiroga foram realizados testes com as diversas técnicas citadas acima, utilizando diferentes bases de dados, e os resultados atingidos em campo (resultados reais) se aproximarem dos resultados previstos no treino.

Calili (Calili, 2005) apresenta uma metodologia para detecção de fraudes em baixa tensão em que se baseia na pesquisa de posses e hábitos de consumo (PPH) na área de actuação da empresa. Essa pesquisa considerou também questões socioeconómicas, como renda familiar, peso da conta de luz no orçamento e localização da residência com o objectivo de classificar os consumidores em normais, anormais e fraudulentos. Foi feita a selecção dos grupos válidos utilizando a base de dados da empresa usando um mapa organizável de Kohonen. Após a selecção dos grupos válidos a classificação de cada grupo quanto a incidência de consumidores normais, anormais e fraudulentos foi feita através de um processo de análise fuzzy. Ainda segundo (Calili, 2005), essa etapa foi realizado a partir das respostas da pesquisa de posses e hábitos. Para a selecção das variáveis relevantes da pesquisa foram aplicados modelos estatísticos nas respostas dos consumidores, gerando curvas e gráficos que foram utilizadas para a selecção das informações mais importantes para analise fuzzy. A referida metodologia apresentou bons resultados na identificação de grupos de consumidores anormais devido a baixa quantidade de respostas obtidas com a Pesquisa de Posses e Hábitos (PPH). As principais desvantagens estão na dificuldade de fazer o PPH para todos os consumidores da empresa e na classificação que é feita por grupo (C*luster*) e dentro desse grupo é necessário identificar quais os clientes que são realmente fraudulentos.

3.8.4. Tendências, Desafios e Perspectivas

De modo geral podemos afirmar que quando o processo de *Data Mining* for aplicado numa empresa, melhora a interacção entre este e os clientes, aumentando as vendas e dirigindo as estratégias de negócio. O *Data Mining*, porém, pode ser aplicado a qualquer conjunto de dados, sejam eles oriundos da medicina, economia, astronomia, geologia, energias, entre outras áreas de estudo.

Observando as aplicações académicas e corporativas acima, é possível citar algumas das fortes tendências da área, bem como os desafios e as perspectivas de *Data Mining*:

- Os casos citados e outros possuem potenciais não explorados. O volume e complexidade dos dados, aliados à peculiaridade das respectivas operações, revelam em cada aplicação um grande conjunto de oportunidades para actividades de pesquisa que aperfeiçoem e inovem métodos e tecnologias (Sarawagi et al. 1998);
- O alto nível de integração de plataformas e de bases de dados remotas demanda uma igual integração das ferramentas de *Data Mining* com diferentes sistemas de bases de dados, *Data Warehouses* e *Web* (Silva & Robin 2002);
- Devido à interactividade das tarefas, linguagens que especifiquem consultas e processos são muito bem vindas em ambientes de KDD, uma vez que a utilização de diferentes ferramentas, o controle do fluxo do processo e gestão do conhecimento demandam esforço extra na ausência dos recursos providos por uma linguagem (Silva & Robin 2004);
- Visual *Data Mining* concerne ao emprego de recursos de computação gráfica (CG) para evidenciar padrões em bases de dados. A evolução de ambas as áreas (KDD e CG) amplia as oportunidades de relevantes trabalhos neste domínio (DBMiner 2000);
- *Data Mining* complexos e semi-estruturados - além das gigantescas bases de dados convencionais, repositórios de dados não convencionais (imagens, textos, graficos, Web, multimédia etc.) apresentam-se como grandes motivações para pesquisas e projectos inovadores (Simoff et al. 2002);

- Protecção de privacidade e segurança de dados - os frequentes ataques a sistemas computacionais, especialmente através da Web, oferecem um excelente campo de aplicação para métodos de *Data Mining* em tempo real como, por exemplo, avaliação de comportamento e padrões de uso.
- *Data Mining* para identificação de fraudes – analisar comportamentos actuais para se detectar fraudes.

4. Estudo de Caso - Modelação do Sistema de Informação Proposto a Detecção de Perdas na Rede Eléctrica

O capítulo quatro trata especificamente da apresentação do **modelo de Sistema de Informação proposto a detecção de perdas na rede eléctrica** baseado na aplicação de tecnologias de extracção de conhecimento. Nele é descrito o modelo fazendo uma descrição detalhada do estudo realizado e dos objectivos atingidos abordando os métodos, técnicas e ferramentas de análise dos dados utilizados para depois apresentar os resultados finais obtidos.

Serão analisadas primeiramente os dados dos postos de transformação para identificação de áreas criticas e transformadores com perdas de energia para em seguida, fazer a apresentação do processo de análise dos dados dos clientes finais pertencentes a estes transformadores utilizando a tecnologia *Data Mining*.

4.1. Descrição do Estudo

O aumento das perdas comerciais em energia eléctrica na cidade da Praia, traduz-se na queda de receita e na perda de energia eléctrica para a empresa. Assim, a busca de alternativas para minimizar as perdas tem uma relevância estratégica para a Electra.

De referir que as perdas comerciais estão relacionadas directamente com as ligações clandestinas que fazem o desvio de energia eléctrica da rede de distribuição directamente para as instalações do cliente, sem passagem pelo contador de energia, falhas na medição ou erros de leitura.

Neste contexto, o sistema proposto tem como **objectivo geral** identificar casos suspeitos de perdas de energia eléctrica analisando, o sistema de distribuição desde a saída dos transformadores até o cliente final. Este sistema permite identificar transformadores e regiões onde ocorrem os maiores desvios de energia, para depois identificar os possíveis clientes fraudulentos pertencentes a estes transformadores e regiões.

Numa primeira fase pretende-se restringir o problema utilizando a tecnologia *OLAP* para a identificação de anéis, zonas e postos de transformação com perdas de energia para posteriormente identificar os consumidores fraudulentos pertencentes a estes postos de transformação utilizando a tecnologia *Data Mining*.

Portanto, como **objectivos específicos** do estudo, apontamos:

- Construir os padrões de informação que apontem as evidências, suspeitas ou indícios de fraudes;
- Identificar e analisar a situação dos postos de transformação, anéis circundantes bem como as zonas geográficas com perdas de energia eléctrica;
- Detectar consumidores finais com irregularidades de energia eléctrica;

Neste contexto, este estudo, tem como pergunta de partida**:**

Quais os padrões de informação existentes na Base de Dados, que permitem a descoberta de indícios, evidências ou, pelo menos, suspeitas da ocorrência de perdas de energia eléctrica?

A Electra, empresa de produção e distribuição de energia eléctrica em Cabo Verde pretende diminuir o percentual de perdas de energia eléctrica existente actualmente

principalmente na Cidade da Praia, visto que é a Cidade com maiores perdas (cerca de 34% de perdas de energia eléctrica).

A cidade da Praia foi escolhida para a aplicação da metodologia de detecção de possíveis perdas de energia eléctrica porque contempla nela diferentes realidades sociais, desde habitações populares de classe baixa, média e alta, a diversos estabelecimentos comerciais, industriais, públicos, estado que correspondem a diferentes tipologias de consumo e perdas.

Os dados adquiridos da empresa para o estudo de caso são referentes aos registos dos clientes, leituras de contadores e postos de transformação, consumo e facturação mensal dos clientes, durante o ano de 2008. Pretende-se analisar esses dados com as ferramentas (*Sql Server 2008*, Excel 2007 e *Weka*) utilizando as tecnologias (*Data Warehouse*, *Olap* e *Data Mining*) para extracção do conhecimento com capacidade de fornecer resultados confiáveis para o estudo das perdas de energia eléctrica.

Para atender as necessidades de análise das informações, foram criadas dois *Data Mart*. Um *Data Mart* a partir dos dados das leituras dos postos de transformação (PT) para identificação de possíveis zonas geográficas e PT´s com perdas de energia eléctrica e um outro a partir dos dados dos clientes finais, como dados de consumo, facturação e contadores para possíveis identificações de clientes fraudulentos.

4.2. Construção da Data Warehouse dos Postos de Transformação

Como mencionado anteriormente na saída dos transformadores são feitas leituras mensais de Potencia em (KVAH) e Intensidade (A) de cada fase. De salientar que não é feito nenhum tratamento sobre esses dados de leituras por parte da empresa, Electra. Esses dados de leituras são recolhidos e depois consultados quando necessários, quando deveriam ser tratados para obtenção de informação, visto que estes dados são importantes para a verificação de possíveis localidades com perdas de energia, para identificar áreas críticas com maior incidência de perdas.

Também permitem identificar Postos de Transformação (PT) em sobrecarga ou seja PT

com potência consumida maior que o normal e poder assim fazer o acompanhamento ou a gestão de carga de cada fase.

Este método permite indicar quais os transformadores que alimentam os clientes da empresa com maior índice de perdas e os transformadores sobrecarregados, permitindo à empresa concentrar seus esforços de fiscalização nesses circuitos.

Para a concepção da base de dados foram feitas recolhas e análise de alguns documentos e leituras mensais dos diferentes Postos de Transformação de diferentes localidades referentes ao ano de 2008 facultados pela Electra. Estes documentos permitiram fazer uma abordagem total sobre as informações recolhidas que posteriormente foram processadas de forma a dar origem ao *Data Mart* dos consumos dos PT´s.

De salientar que foram recolhidos dados da intensidade de cada fase, as potencias activas (KWH) e reactivas (KVAH) de cada transformador pertencente a uma determinada zona ou local de consumo referentes ao ano 2008.

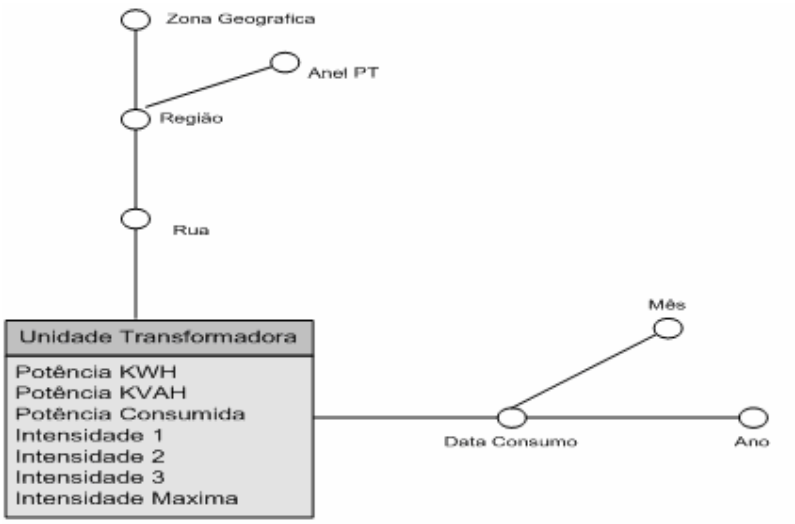
4.2.1. Pré-Processamento e Transformação dos Dados

Esta etapa permitiu detectar os erros e inconsistência no registo das leituras dos postos de transformação. Foi efectuada esta etapa no sentido de fazer uma limpeza dos dados afim de adequar e carregar os dados necessários no *Data Warehouse*.

Alguns registos das leituras encontrados nas fichas de leituras em Excel foram excluídos por não apresentarem informações precisas ou seja, existiam dados incompletos referentes aos valores das leituras dos consumos mensais dos PT´s, alguns dados das leituras das intensidades também estavam incompletos e outros duplicados, tendo estes últimos sido eliminados. Foram encontrados dados dos valores dos consumos e intensidades de cada PT modificados e datas trocadas. Foi fácil a identificação desses dados incompletos na medida em que o consumo e a intensidade de cada PT activo têm que oscilar sempre no sentido positivo.

4.2.2. Modelação Dimensional

A etapa seguinte consistiu no desenho do modelo de dados multidimensional e na criação do respectivo *Data Mart*. Buscou-se abordar o assunto referente as leituras e consumos dos postos de transformação pertencentes as localidades e anéis diferentes. A figura seguinte ilustra o modelo de dados proposto usando a notação proposta em (John Wang, 2008).



Esquema 9 - Data Warehouse dos Postos de Transformação

Neste modelo consideramos as dimensões geográficas e temporal. A dimensão geográfica identifica o local de consumo, a zona, a região e o anel geográfico pertencente aos postos de transformação enquanto que a dimensão temporal identifica as datas dos registos de consumo nos referidos postos.

A tabela de factos (Unidade Transformadora) contém os dados identificadores de "movimento" ou seja as medidas. No caso real do nosso sistema temos o valor dos consumos dos PT´s em KVAH, mas para podermos identificar perdas e detectar PT´s sobrecarregados temos que ter outros dados. Através do atributo POTENCIA_PT_KVAH podemos facilmente chegar aos outros atributos calculando as fórmulas seguintes:

$$S = P.\cos(\varphi)$$

Onde, **S** representa a potência em KWH, **P** potência em KVAH e $\cos(\varphi)$ é 0.8. E assim

temos o valor do atributo POTENCIA_PT_KWH.

Para calcular o valor do atributo da potencia consumida de todos os clientes pertencentes a um transformador:

$$S = \sqrt{3}.U.I \;/\; 1000$$

Em que, **S** é a potência em KWH, **U** é a tensão do transformador (400 V) e **I,** média das intensidade entre as fases, que são as intensidades INT_F1, INT_F2, INT_F3.

Estes atributos mencionados acima permite-nos calcular o percentual de folgas existentes em cada um dos postos de transformação e assim definir a situação de cada um dos PT´s. Um Posto de Transformação pode estar numa situação crítica ou sobrecarregada se a sua folga percentual for negativa. Se a sua situação for de atenção especial quer dizer que o Posto de Transformação está em risco de perdas ou sobrecarregados de potência. Situação normal quer dizer que o Posto de Transformação está com um percentual de folga aceitável, nos parâmetros normais.

A fórmula para calcular a fase máxima é a seguinte:

$$\text{Max (U)} = (\text{Max (I)} / P / \sqrt{3}*U) / 1000$$

Em que, Max (I) é o máximo das intensidades de cada fase a dividir por um calculo predefinido pelos técnicos. Em que o P é a potencia em KVAH e U a tensão de (400 V). Para calcular a folga é necessário fazer o calculo em baixo.

$$\text{Folga \%} = (1 - \text{Max (U)}) * 100$$

O processamento analítico dos dados foi realizado usando o Microsoft Excel 2007 que possui várias ferramentas cuja finalidade é facilitar a análise e visualização multidimensional dos dados sob a forma de tabelas ou gráficos, fornecendo desta forma ao analista informações detalhadas dos postos de transformação e zonas com mais perdas.

Esta metodologia permitiu procurar respostas para um conjunto de questões pertinentes, das quais destacamos as seguintes:

- Verificar e analisar quais foram as zonas geográficas reincidentes em perdas após fiscalização;
- Identificar Postos de Transformação e Anéis com maior número de perdas durante o ano 2008;
- Identificar Postos de transformação em sobrecargas de potência (desfasamento de intensidade);

- Identificar os períodos do ano com maior índice de perdas por anéis e zonas geográficas (zonas em alerta vermelho ou seja, as zonas críticas)
- Identificar zonas e anéis de acção urgentes (zonas em alerta amarela ou atenção)
- Analisar a evolução do consumo (KWH) por zonas geográficas e PT's durante o ano 2008;
- Analisar a variação da curva percentual (folga) de cada posto de transformação durante o período de 2008;
- Situação anual das zonas em relação ao consumo durante o ano 2008.

Os resultados destas questões podem mostrar as zonas geográficas com maiores probabilidades de incidências de fraudes de energia eléctrica e permite dizer quais os transformadores que alimentam os clientes com maiores índices de perdas comerciais e que estão sobrecarregados auxiliando assim as tomadas de decisão a realizar.

Para cada uma das questões listadas acima foram gerados relatórios e gráficos para análise dos dados a fim de encontrar informações úteis. A seguir serão apresentados os resultados obtidos.

Primeira Questão

A primeira consulta realizada envolve a verificação e análise das zonas geográficas reincidentes de perdas após fiscalização durante o ano 2008. De referir que as zonas reincidentes em perdas de energia eléctrica são as zonas situadas no anel da fazenda I como Eugénio Lima, Castelão e Fazenda I. Estas zonas foram identificadas como reincidentes de perdas através da análise da situação durante o ano 2008 em comparação com os dados do ano anterior (2007). As medidas importantes que permitiram chegar a estas conclusões foram as intensidades e potências em KWH, calculando e definindo posteriormente como atributos as fases máximas e média e as folgas em percentagem.

O Gráfico 1 apresenta o caso real da zona de Eugénio Lima pertencente ao anel da Fazenda I que iniciou o ano 2008 mais precisamente o mês de Janeiro com a sua situação de alerta amarela.

Esta região requeria de forma urgente acções de fiscalização na zona e nesse respectivo posto de transformação (PT). Não houve acções de prevenção e os meses subsequentes foram críticos, com cortes frequentes de energia eléctrica principalmente nas horas de ponta, porque o valor consumido pelos clientes no referido PT que alimenta a zona

ultrapassou o valor do consumo total do PT.

A partir do mês de Maio começaram alguns trabalhos de fiscalização mas foi a partir do mês de Setembro que a situação melhorou com o aumento da potência do PT em KVAH e acções mais severas de fiscalização. De realçar que foi detectado que o PT de Eugénio Lima estava sobrecarregado de consumo depois de sucessivos disparos.

Este processo foi usado para detectar outras situações anómalas, como a reincidência de fraudes e postos de transformação sobrecarregados, fazendo comparações, cruzamentos e análise dos dados das leituras e visualização dos mesmos em diferentes perspectivas.

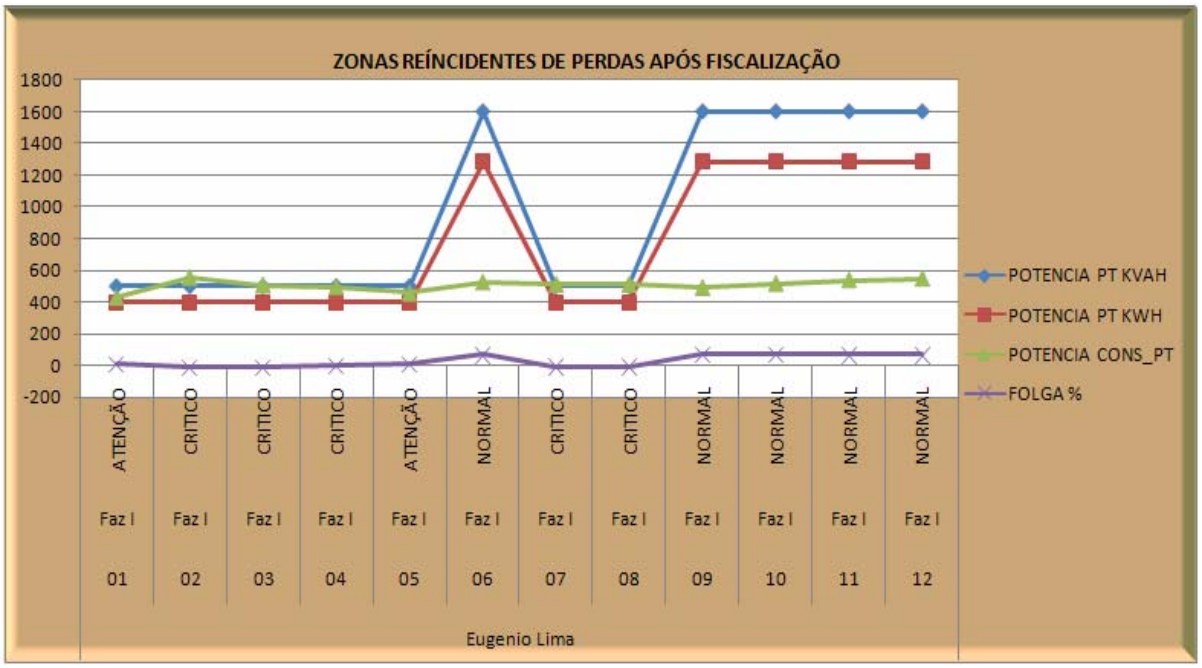


Gráfico 1 - Gráficos das Zonas Reincidentes de Perdas

Segunda Questão

A segunda questão permite identificar os postos de transformação, referidas zonas geográficas e os seus anéis circundantes com maior número de perdas durante o 1º trimestre do ano 2008. Como observado no gráfico 2 o valor do consumo dos clientes PT nas zonas seleccionadas ultrapassa o valor da potência do PT o que na óptica dos técnicos especialistas corresponde a perdas comerciais avultadas para a empresa. Conseguiu identificar os postos de transformação, zonas e anéis ou saídas circundantes com maior número de perdas durante o 1º trimestre do ano de 2008. Com o cruzamento das informações das medidas e dimensões conseguiu-se num certo período de tempo definir as

zonas e PT´s com possíveis perdas de energia eléctrica analisando graficamente o percentual das folgas comparando com as potencias de cada PT e a potencia consumida pelos clientes.

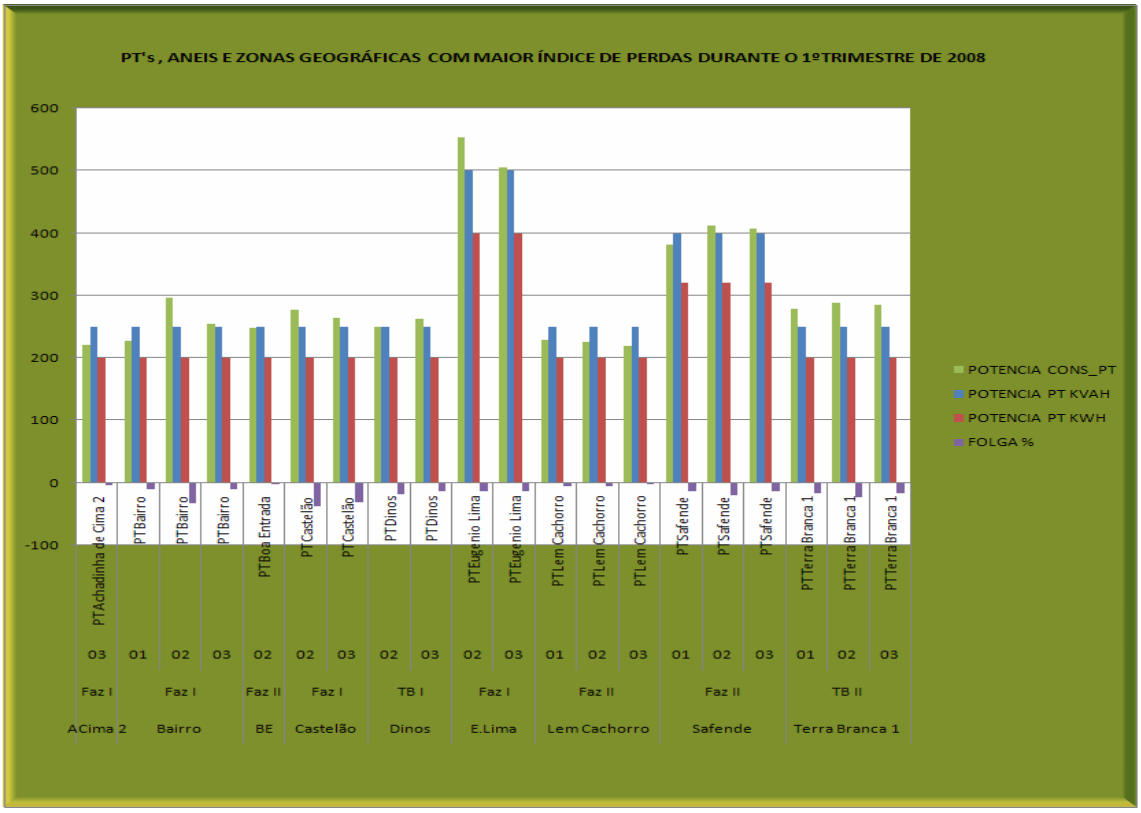


Gráfico 2 - PT's, Aneis e Zonas com maior índice de Perdas

Terceira Questão

Identificar Postos de transformação em sobrecarga de potência durante o primeiro trimestre de 2008 (desfasamento de intensidade). O desfasamento entre as intensidades é obtido através da diferença entre as mesmas. O valor de uma intensidade não pode ser muito superior a outra, caso for, verifica-se que existe perdas e sobrecargas nesta fase e selecciona os clientes ligados à esta fase para inspecção.

Em termos técnicos podemos verificar que um posto de transformação pertencente a um anel está sobrecarregado somente pelas intensidades. Se o desfasamento entre as intensidades for muito grande verifica-se que há perdas nesta fase eléctrica. Pode-se também verificar que um PT está sobrecarregado se a potencia PT KWH for inferior à

Potência Consumida.

O gráfico 3 representa os postos de transformação e as respectivas zonas com sobrecargas de potência e intensidade, portanto possíveis perdas de energia durante o 1º trimestre de 2008.

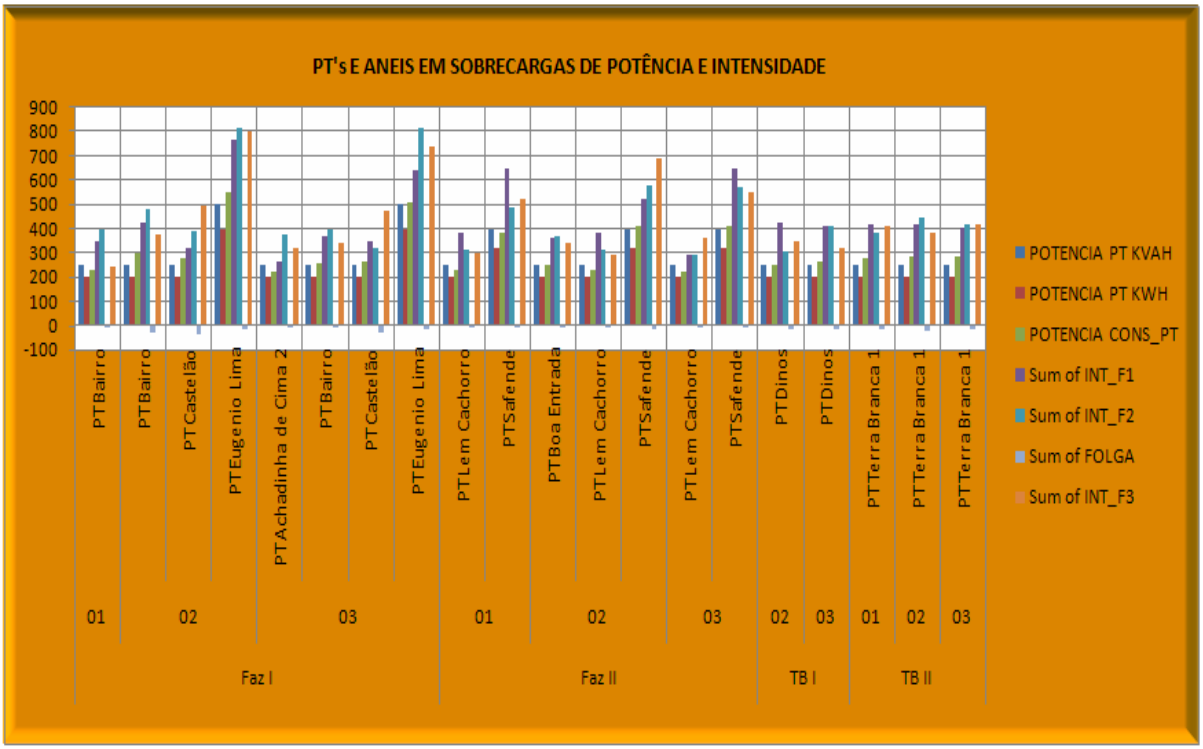


Gráfico 3 - PT´s e Aneis em sobrecargas de potencia e intensidade

Quarta Questão

Períodos do ano com maior índice de perdas por anéis e zonas geográficas (zonas em alerta vermelha ou seja zonas criticas). Para a identificação das zonas, postos de transformação e os respectivos períodos com maiores índices de perdas faz-se uma comparação entre a potência de cada PT com o percentual das folgas e obtêm-se a situação critica anual de cada PT.

Uma análise global dos dados permitiu, concluir que os meses de Fevereiro, Março, Outubro, Novembro e Dezembro registam maiores perdas e que as zonas mais críticas são Eugénio Lima, Castelão, Achadinha de Baixo e Calabaceira.

O gráfico 4 apresenta os meses com maior índice de perdas durante o ano de 2008 no anel da fazenda – I. Estas zonas estão em alerta vermelha com sucessivas quedas de tensão e cortes devido ao aumento de ligações clandestinas nesse período.

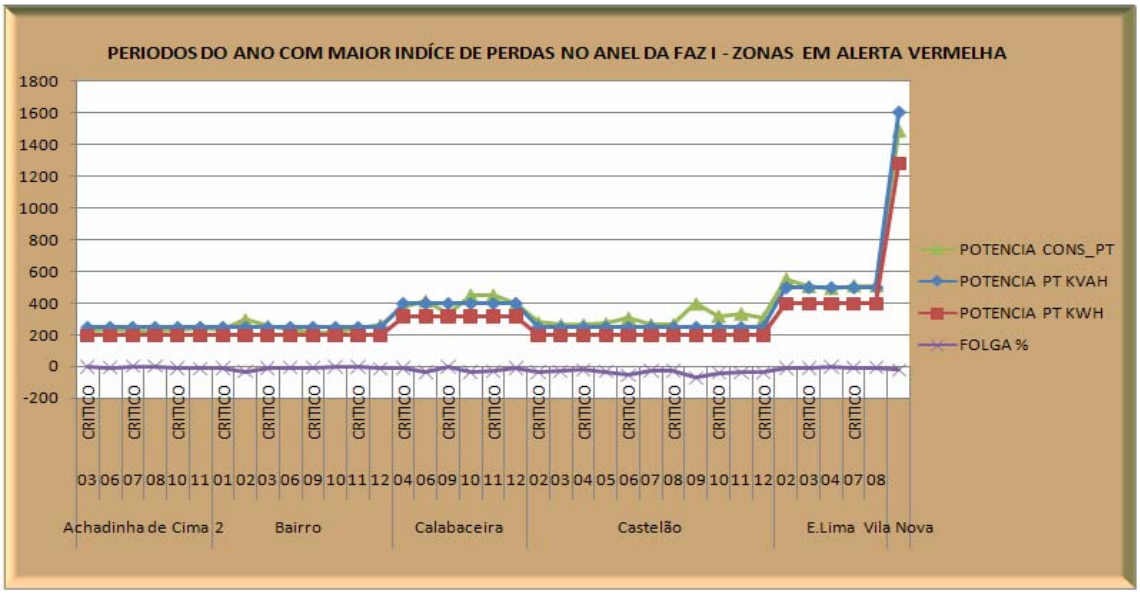


Gráfico 4 - Períodos do ano com maior índice de perdas no anel da Fazenda 1

Quinta Questão

Identificar zonas e anéis de acção urgentes, zonas em alerta amarelo ou atenção durante o primeiro trimestre de 2008. As zonas em alerta amarelo durante o primeiro trimestre são de maioria pertencentes aos anéis da Fazenda I (Achadinha de Cima I, Achadinha Cima II) e II (Paiol, São Filipe, Ponta d´água) e da Terra Branca I (Dinôs e Palmarejo 5). Se o calculo da fase máxima de cada um dos PT´s for menor que 1 a situação é classificada de atenção.

É aqui que os técnicos especialistas têm que combater as perdas nas zonas. O gráfico 5 mostra que o limite percentual permitido para as folgas nos postos de transformação (20%) está quase consumido. Neste contexto podemos frisar que se as perdas não forem combatidas nesta fase, vão aumentar conduzindo a uma situação critica no futuro.

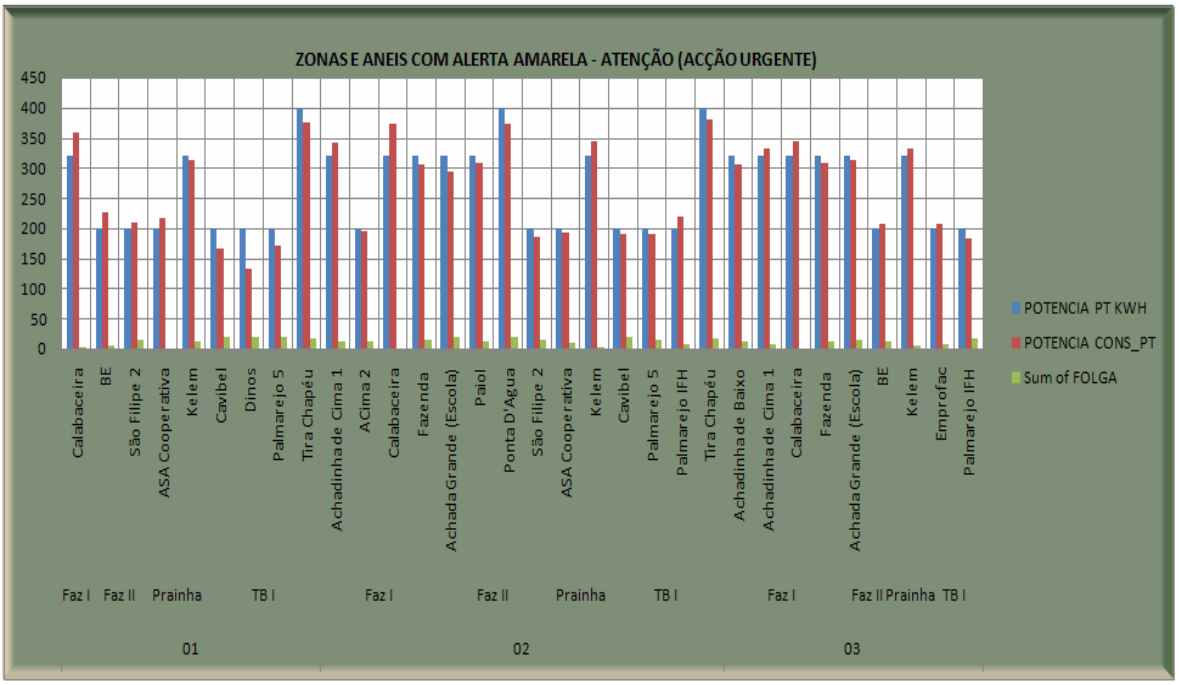


Gráfico 5 - Zonas e Anéis com alerta Amarela

Sexta questão

Analisar a evolução do consumo (KWH) por zonas geográficas e PT's durante o ano 2008. O gráfico apresenta a evolução do consumo dos clientes do PT do Palmarejo 3 e 4 pertencentes ao anel da Terra Branca I, em que a situação parece normal durante o ano com o percentual das folgas sempre acima do valor desejado excepto a zona do Palmarejo 4 que teve uma situação crítica no mês de Setembro fazendo com que o consumo ultrapassa-se o limite da potencia em KWH.

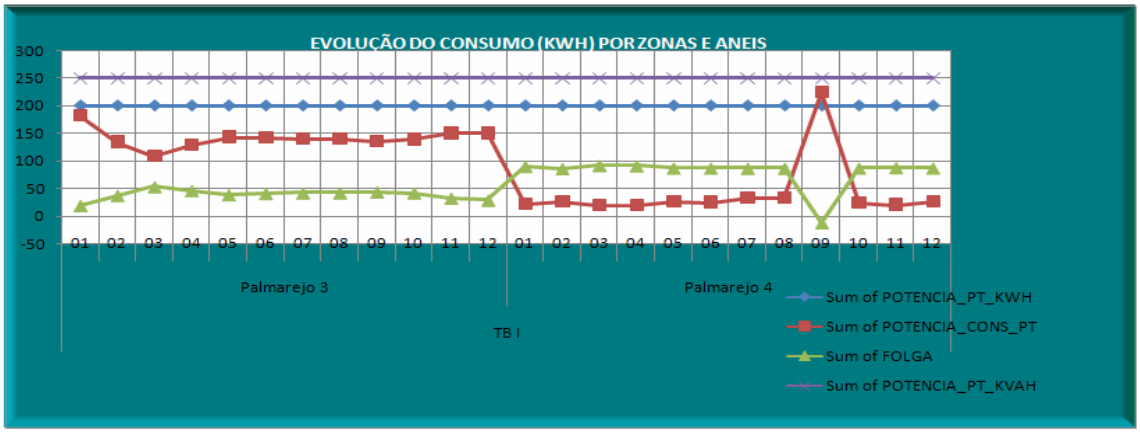


Gráfico 6 - Evolução do Consumo em KWH por zonas e anéis Palmarejo

Sétima questão

Analisar a variação do valor da folga em cada posto de transformação durante o período de 2008. Esta análise contribui para a identificação das zonas com maior variação de consumo e folgas dos PT´s durante o ano permitindo a verificação dos meses com maior variação de consumo e possíveis perdas.

Os gráficos abaixo fazem a comparação entre dois postos de transformação de dois anéis diferentes em que a curva percentual das folgas durante o ano de 2008 esteve praticamente negativa e que no outro caso esteve positiva.

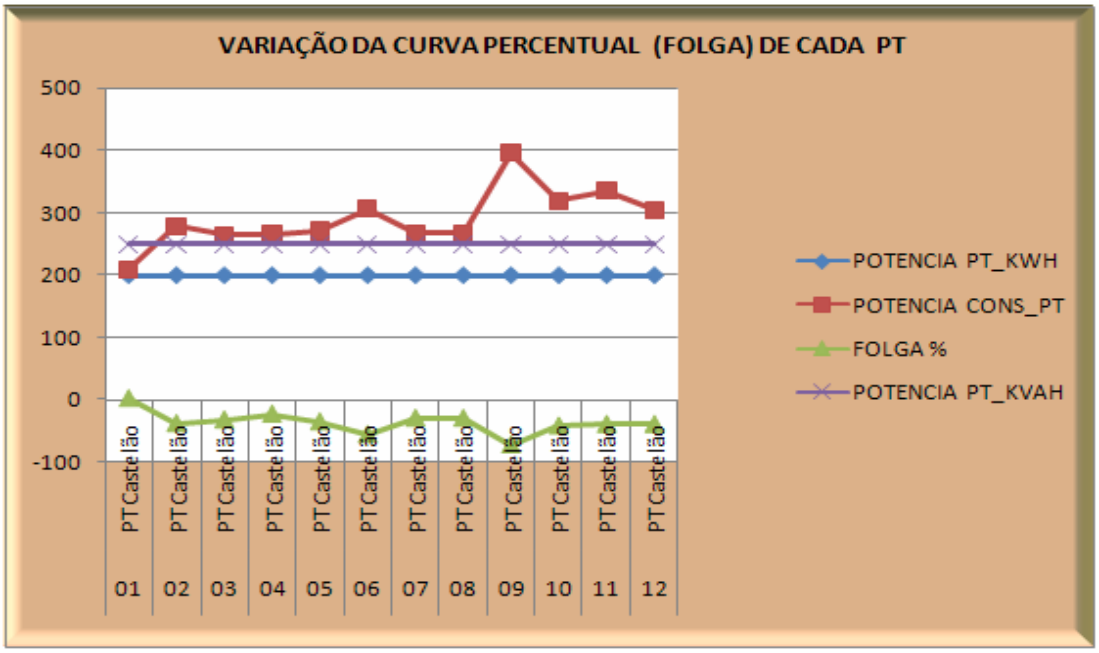


Gráfico 7 - Variação da curva percentual em cada PT – Castelão

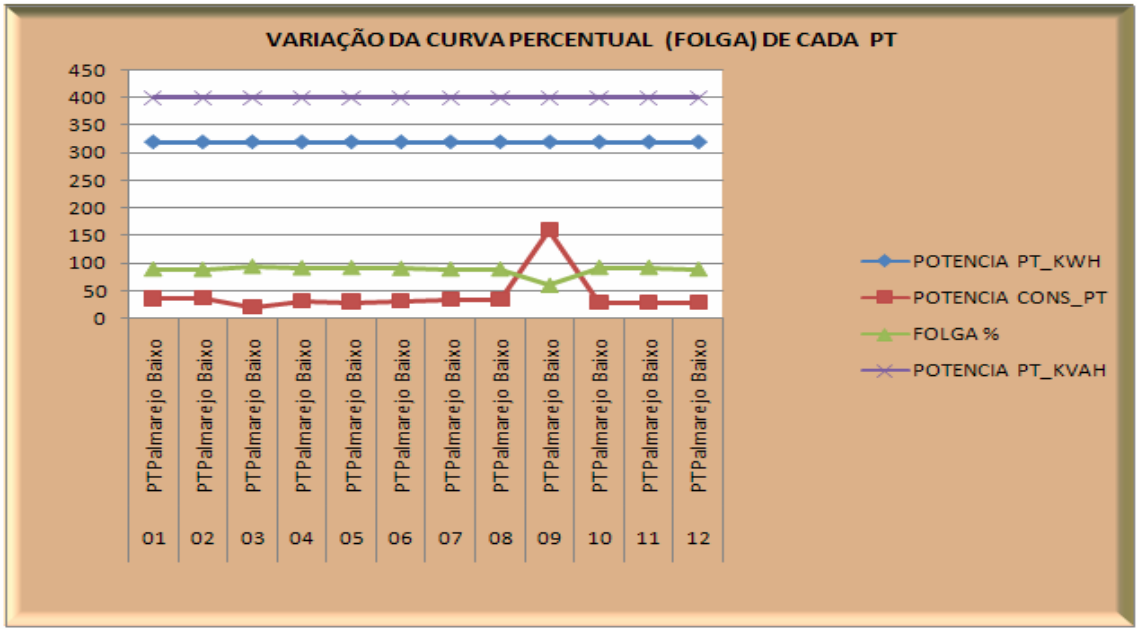


Gráfico 8 - Variação da curva percentual em cada PT - Palmarejo

Oitava questão

Com estes gráficos podemos analisar a situação anual das zonas em relação ao consumo durante o ano 2008. De referir que um Posto de Transformação encontra-se numa situação critica quando a fase média é maior que 0.8 (valor atribuído pelos técnicos). Sendo a fase média calculada como a média das intensidades a dividir por √3*0.4. A classificação de uma zona ou um PT como sendo critica mostra-nos que este PT pode estar sobrecarregada e com possíveis perdas de energia. Se a situação for de atenção, o calculo da fase máxima de cada um dos PT´s tem que ser menor que 1, enquanto que se a situação for normal quer dizer que o Posto de Transformação está com um percentual de folga aceitável, nos parâmetros normais.

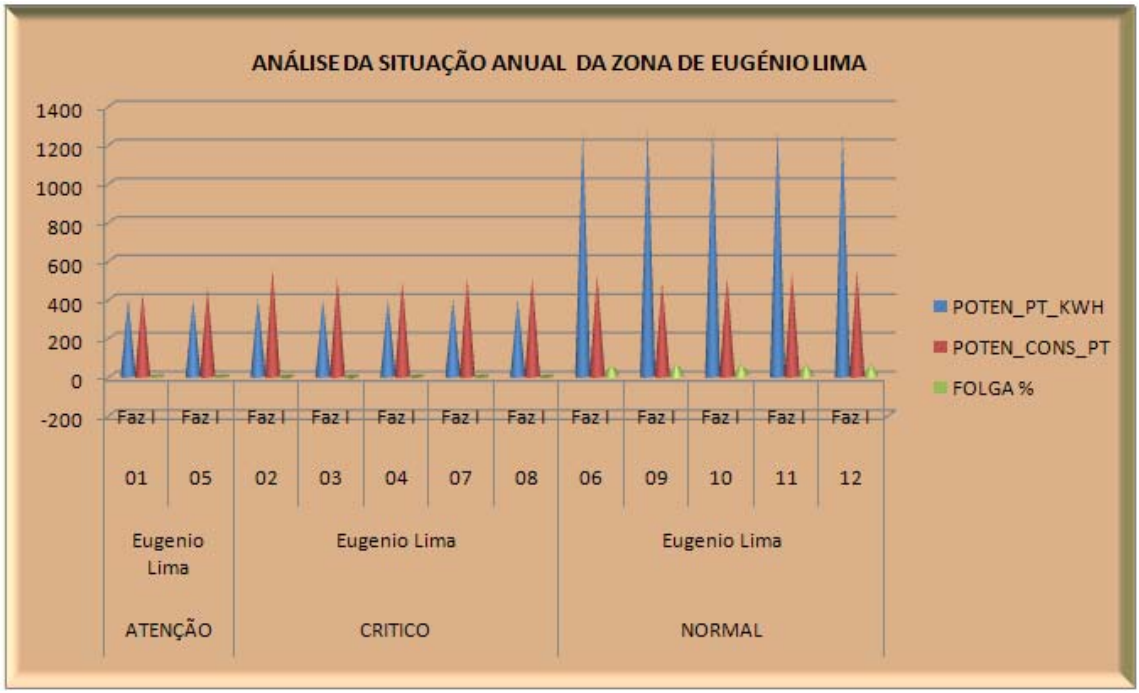


Gráfico 9 - Análise da situação actual da zona Eugénio Lima

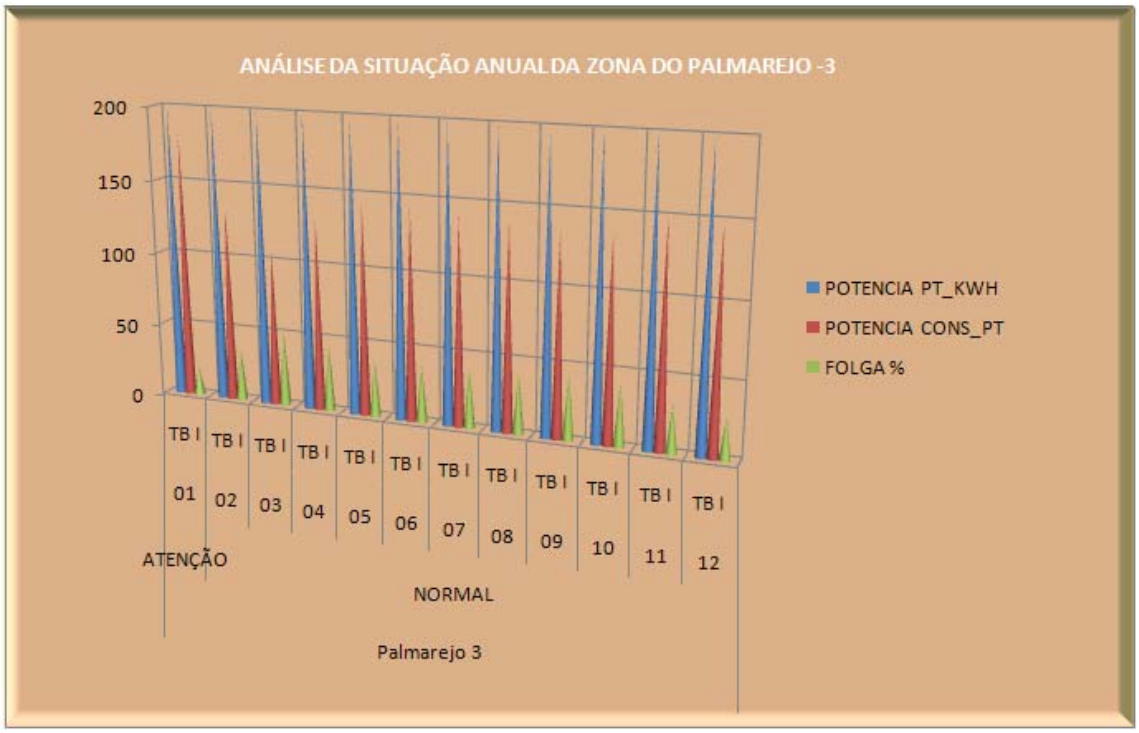


Gráfico 10 - Análise da situação actual da zona Palmarejo

4.2.3. Análise Final dos Resultados

Após a extracção e análise dos dados dos Postos de Transformação podemos dizer que a *Data Warehouse* implementada já possui um *Data Mart* das Unidades de Consumos dos Transformadores completamente funcional e pode ser analisada a qualquer momento com as ferramentas OLAP para a consulta e geração de relatórios. O recurso a ferramentas *OLAP* permitiu-nos ter uma visão multidimensional dos dados e definir as premissas a serem utilizadas no processo de *Data Mining* realizado a seguir, das quais destacamos:

- Estão identificadas os anéis, as zonas geográficas e os postos de transformação que têm mais perdas de energia eléctrica,
- Estão reunidas as informações de consumo e carga mensal (KWH) de cada fase dos transformadores de cada zona.
- É possível identificar postos de Transformação que estão numa situação de alerta, que tenham perdas de energia eléctrica ou que estão sobrecarregados.

Neste âmbito estão **identificadas as áreas críticas** com maior incidência de perdas comerciais dentro da área de concessão da empresa, factor essencial ao combate ao problema, fazendo a comparação entre a energia consumida à saída dos transformadores e a energia consumida pelos clientes de cada região pertencente a este transformador.

Com os resultados dos dados dos PT's pode-se fazer o **balanço energético,** diferença entre a energia medida pelos contadores instalados junto aos postos de transformação e a energia medida pelos contadores instalados nas unidades consumidoras ligadas aos referidos transformadores. Concentrando todas as informações do consumo dos transformadores na base de dados, é possível apontar facilmente a diferença entre energia distribuída, menos o consumo medido. A diferença restante é uma aproximação do valor das perdas comerciais, ou fraudes. Este método permite dizer quais os transformadores que alimentam os consumidores com maior índice de perdas comerciais, permitindo à empresa concentrar seus esforços de fiscalização nesses circuitos.

De realçar que foram aplicadas varias funções de agregação aos dados da *Data Mart* como o cálculo de médias, somas, mínimos e funções gráficas que auxiliaram no processo de identificação de perdas nos transformadores

A partir deste momento podemos iniciar o modelo de investigação de detecção de suspeitas de fraudes nos consumidores finais, baseado na aplicação da tecnologia de extracção de conhecimento *Data Mining* para análise dos dados relativos aos registos dos clientes.

4.3. Construção da Data Warehouse dos Clientes Finais

O sistema tem como objectivo apoiar na identificação de possíveis clientes fraudulentos através da aplicação de regras que identificam suspeitos de fraudes a partir do histórico de consumo e outros dados disponíveis.

Pretende-se que este sistema defina e analise a série histórica anual do consumo e facturação do consumidor para identificar a velocidade mensal de variação dos indicadores.

Podemos desde já identificar alguns dos indicadores básicos que podem ser útil para identificação de possíveis fraudes:

- Estrutura do consumo – histórico e sua evolução no tempo;
- Consumo por consumidor residencial – histórico e sua evolução no tempo, avaliando os componentes que o determinam;
- Correlações entre os consumos, valor dos consumos e das categorias de consumo.

A estrutura ou o **histórico de consumo** consiste na comparação de consumos trimestrais e semestrais de um determinado consumidor que não oscila de forma brusca quando comparados em um curto período de meses. Esta regra avalia se o consumo do último mês reduziu um certo percentual em relação ao consumo dos últimos três meses. Caso tenha reduzido existirá a possibilidade de terem ocorrido anomalias de leitura ou facturação nesse período.

Pode-se ainda classificar os **clientes como confiáveis e não confiáveis** com base no pagamento da factura e o valor da conta (factura a ser paga). Por exemplo, é possível identificar os clientes com pagamentos em atraso ou intervalos de consumos e contas aceitáveis para uma determinada categoria de consumo.

Através destes indicadores podem ser seleccionados clientes segundo regras a serem

aplicadas sobre o conjunto de dados de investigação, comparando resultados individuais de cada regra definida.

Definidos os indicadores, vamos descrever o processo de preparação da base de dados para a extracção do conhecimento. De realçar que a base de dados da Electra não está preparada para utilização de *Data Mining*, existe uma longa caminhada até que os dados transaccionais sejam transformados e armazenados (ETL) em uma *Data Warehouse* para interligar dados de forma centralizada disponibilizando-os para processos de extracção do conhecimento.

Para isso foram processados os registos dos consumidores da base de dados fornecida pela Electra, relativos a algumas zonas da cidade da praia do ano de 2008.

De salientar que a base de dados da Electra, existente é implementada em *Oracle* com suporte em uma empresa Portuguesa, o que dificulta a extracção e relacionamento dos dados das tabelas requeridas.

4.3.1. Pré-Processamento e Transformação dos dados dos Consumidores

A principal informação utilizada é o histórico de consumo de cada unidade consumidora. Os técnicos indicaram que os dados de consumo num período de seis a doze meses seriam suficientes para fazer as análises.

Além dos consumos, foram requisitadas também os históricos de anomalias de leituras. Essas anomalias de leitura são registadas no momento em que o leitor realiza o registo do consumo a partir da leitura do contador de cada unidade consumidora.

De salientar que algumas dessas anomalias possuem informações que podem invalidar o consumo do mês, e portanto são importantes para o uso no sistema. Outra informação similar às anomalias de leitura são as anomalias de facturação que são registadas automaticamente pelo sistema e indicam que houve uma anomalia no mês em observação da qual destacamos: tarifário inexistente, escalão de consumo inválido, aparelhos inexistentes, aparelho duplicado no local, etc.

Por exemplo os dados das anomalias eram gerados em forma de códigos, era necessário verificar as descrições destes códigos muitas vezes em tabelas diferentes, para depois transforma-las em dados de leitura pela *Data Warehouse.*

Foi necessário realizar algumas tarefas para o tratamento dos dados extraídos. A

primeira foi a limpeza dos dados com registos incompletos, com valores dos consumos fora dos limites estabelecidos. Alguns registos também foram excluídos por não apresentarem informações concisas nomeadamente os registos que possuíam o valor de consumo, leitura e contas negativas.

Houve outros casos de limpeza e transformação dos dados em que registos apresentados em forma de códigos, como as siglas, foram transformados em dados significativos. Para a transformação foram necessárias consultar as descrições textuais dos referidos códigos e números. Por exemplo as anomalias de facturação e de leituras eram escritas em forma de códigos pelos técnicos de leituras para evitar problemas com os clientes. As categorias de consumo de cada cliente encontravam-se também no formato de códigos e número e foram transformados em textos descritivos.

Um outro tipo de tratamento dos dados necessários foi a conversão do tipo dos dados. Os registos extraídos no formato texto continha dados de outros tipos, como valores monetários (*Currency*), números inteiros, reais e datas.

Todas as ocorrências apresentadas e que foram ajustadas na fase de pré-processamento, configuravam como erros no registo na base de dados da empresa e por isso tiveram que ser inaceitáveis no *Data Warehouse*, pois caso não fossem realizadas limpezas necessárias, iriam provocar erros de semântica nas informações a serem analisadas com *Data Mining*.

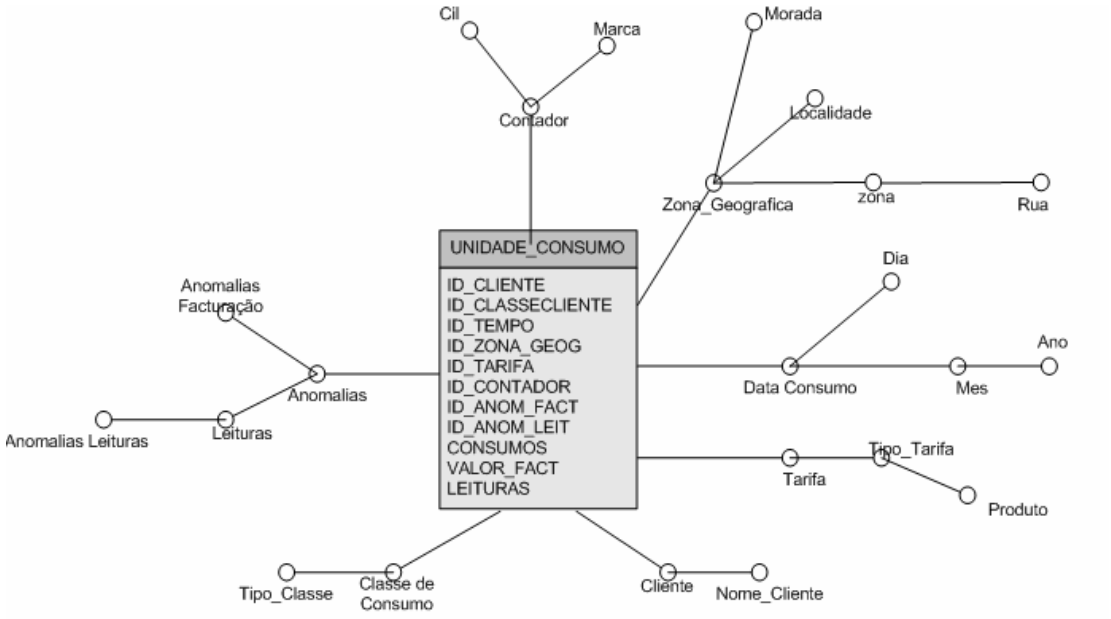
4.3.2. Modelação dos Dados dos Consumidores Finais

De realçar que foram feitas análises do modelo de dados da aplicação de gestão de consumo e facturação (Elag) e da base de dados a fim de seleccionar um subconjunto de dados que permitam criar uma *Data Staging Área*, contendo dados realmente necessários para a criação da *Data Mart* ou *Data Warehouse* de Unidade de consumo. Todo o armazenamento de dados está sendo feita no *Microsoft SQL Server 2008*.

Após a análise do modelo de dados foi definida a granularidade da tabela de factos. Os registos na tabela de factos representam os valores das leituras e consumos dos contadores e clientes respectivamente. Um aspecto importante na identificação de um facto (transacção) é a sua característica evolutiva, ou seja, mudança de características ao longo do tempo, podendo ser sempre questionado sobre sua evolução ao longo de um período.

Foi utilizada o modelo em estrela e definida uma tabela de factos, Unidade_Consumo responsável pelo armazenamento das medidas numéricas do negócio como valor consumido por cada consumidor, valores das leituras de cada contador fornecidas pelo cliente final e pelas verificações, e o valor da facturação.

A figura abaixo apresenta o modelo dimensional do sistema de Unidades Consumidoras para possível detecção de suspeitas de fraudes de energia eléctrica.



Esquema 10 - Data Warehouse Unidade de Consumo

Basicamente o conjunto dos dados definidos no modelo dimensional apresentado divide-se em três grupos de informações:

- Informações dos registos dos consumidores.
- Informações históricas de consumo do consumidor.
- Informações sobre as anomalias de facturação e leituras dos contadores.

O registo das informações gerais sobre cada unidade consumidora inclui campos como a dimensão cliente (Dim_Ciente) que descreve os dados de identificação do cliente, tipo de clientes (Dim_TipoClientes) que descreve os tipos de clientes (domésticos, autarquias, instituições, estado tesouro, empresa pública, comércio, indústria ou agricultura).

A dimensão zona geográfica (Dim_ZonaGeografica) descreve as localidades das unidades de consumo, a dimensão tempo (Dim_Tempo) descreve o período de facturação.

De realçar que o período de facturação é o mesmo que o período das leituras feitas nos contadores efectuadas pelos leitores. A dimensão tarifa (Dim_Tarifa) descreve o tipo de tarifa que o cliente contratou ou seja, baixa tensão especial (empresas), baixa tensão normal (domésticos) e média tensão enquanto a dimensão (Dim_TipoPotencia) descreve os valores do tipo de potência contratada, isto é, se é monofásica ou trifásica.

Já a dimensão contador (Dim_Contador) identifica os contadores que servem as unidades consumidoras que tem em seu poder a chave de identificação da base que é o CIL – Código de Identificação Local. A dimensão anomalias (Dim_Anomalias) identifica as anomalias de leitura verificadas nos contadores (Ponteiro do aparelho deslocado, display apagado, vidro do aparelho quebrado, etc) e a dimensão tipo de anomalias (Dim_TipoAnomalias) descreve o tipo de anomalias de leitura existente (Anomalia ao nível do aparelho, ao nível do ponto de medida e ao nível do prédio).

Praticamente todas as dimensões, tabelas de factos e métricas serviram para a criação de consultas como:

- Classificação de consumidores quanto ao risco de irregularidades de pagamento ou seja se um cliente é confiável ou não junto da empresa. O valor da conta do consumo de energia é menor que "X", cujo "X" representa o valor mínimo de consumo estimado para um tipo de tarifa para uma determinada classe de consumo.
- Classificação do consumidor quanto a suspeitas de fraudes, calculando a média do consumo dos últimos três meses, ou seja, calculando o mínimo esperado como sendo inferior a 0,6 (variável disponibilizado pelos técnicos da empresa) vezes a média dos últimos três meses.
- Classificação dos consumidores quanto a variação brusca do consumo entre dois meses consecutivos. Uma consulta que determina a média semestral do consumo de cada consumidor.
- Classificação do perfil dos clientes quanto as anomalias de leitura e facturação. Verificação e diagnóstico dos clientes com diferentes tipos de anomalias nomeadamente contador parado, funções congeladas, aparelho inexistente, etc.

4.4. Processo de Extracção do Conhecimento

Pretende-se que o algoritmo utilizado para a extracção do conhecimento análise os dados a fim de produzir uma quantidade de padrões úteis, validos e de fácil entendimento gerando resultados importantes para a tomada de decisões sobre as possíveis perdas encontradas.

Na etapa anterior, usando apenas os dados dos transformadores e a tecnologia *OLAP*, foram detectados os transformadores e possíveis zonas geográficas com mais perdas de energia eléctrica. Pretende-se então, que com a utilização da tarefa de classificação por árvore de decisão detectar os tipos de perdas e irregularidades mais frequentes nas diversas zonas geográficas potencialmente mais fraudulentas e assim determinar o perfil do consumidor final.

A escolha da tecnologia *Data Mining* e da tarefa de classificação por árvore de decisão para possível detecção de perdas de energia surgiu na medida em que árvores de decisão expressam uma forma simples de lógica condicional, simplesmente dividindo tabelas em tabelas menores, seleccionando esses subconjuntos baseados em valores para um determinado atributo. Essas tabelas são apresentadas em forma de árvore, sendo que em seus ramos são apresentadas as perguntas de classificação e em suas folhas estão apresentadas as partições da tabela original ou seja a regra de classificação terá sempre no seu consequente uma resposta ao facto das condições satisfazerem ou não a uma determinada classe previamente definida.

A construção da árvore de decisão objectiva predizer informações, gerando estimativas sobre os consumidores com risco de irregularidades e identificar padrões de comportamento.

A detecção de possíveis perdas de energia bem como a sua prevenção configura-se um problema complexo. Mesmo que o histórico de consumo e o perfil do comportamento do consumidor apresentem claros indícios de fraude é importante que uma segunda investigação seja realizada. Portanto, as informações geradas pelos sistemas de apoio decisão também precisam ser compatibilizadas com outras variáveis do sistema, para que cobranças erróneas não sejam aplicadas aos clientes.

4.4.1. Data Mining Aplicado ao Estudo de Caso

Para o início do processo de *Data Mining* foram realizadas a extracção e o carregamento dos dados a partir da *Data Warehouse*. De realçar que nesta etapa antes de utilizar os algoritmos de *Data Mining* o software *Weka* permite configurar os dados através de varias funções de filtragens, entre os quais, junções, adições, conversão de tipos e formatos, etc.

Neste processo pretende-se analisar características como o histórico de consumo, estado da facturação, zona geográfica residencial e as anomalias de leitura e facturação.

Propõe-se um modelo de extracção do conhecimento que avalia essas características e a partir delas são criadas regras que serão aplicadas sobre os dados dos clientes finais pertencentes a cada um dos transformadores ou anel verificadas anteriormente.

Os atributos da base de dados dispostos para o estudo de caso correspondem ao período de 12 meses. Neste âmbito, uma das principais características a ser analisada primeiramente pelas regras é o histórico de consumo do consumidor. O valor de consumo mensal é comparado com a média do consumo do consumidor que está sendo analisado.

Neste contexto e com o objectivo de ter maior precisão na análise dos dados do histórico do consumo foram criadas três regras que avaliam a média do consumo em períodos diferentes, analisando se o consumo mensal está próximo da média. As regras que avaliam o histórico de consumo são:

- Consumo abaixo do mínimo esperado.
- Consumo superior ao esperado.
- Variação brusca entre dois meses consecutivos.
- Confiança dos clientes

Essas regras são processadas de acordo com um parâmetro de referência que indica qual a variação máxima esperada entre a média e o consumo. Quando o parâmetro é excedido, o consumidor é classificado como anormal.

Pode-se considerar o consumo mínimo por mês de um consumidor como uma regra tendo em vista que ele consome energia. A partir deste facto, consumidores com vários meses com o **consumo abaixo do mínimo esperado** é considerado irregular ou suspeito de possíveis fraudes de energia eléctrica. De salientar que um consumidor pode ter num mês um consumo mínimo devido a ausência do mesmo na sua habitação. Neste contexto pode-se calcular o consumo mínimo esperado como sendo inferior á 0,6 vezes média dos últimos

três meses. Nesse caso verifica-se se ocorreram anomalias de leitura e facturação nesse período. Em caso afirmativo, classifica-se o risco de fraudes como baixo ou médio. Senão, o cliente é seleccionado para inspecção com alto risco de fraudes.

$$\mathbf{CME < 0{,}6* \frac{x1+x2+x3}{N}}$$

Os consumidores também são considerados numa situação irregular caso tenham um **consumo superior ao esperado**. O consumo de um determinado mês não deve variar muito em relação aos meses anteriores, espera-se que o consumo não tenha uma variação brusca superior a três vezes a média dos últimos três meses consecutivos. Se isso acontecer, verifica-se o estado do contador e as respectivas anomalias.

$$\mathbf{CSE >= 3* \frac{x1+x2+x3}{N}}$$

A regra **variação brusca entre dois meses consecutivos** considera que o consumo de um determinado mês não deve variar muito em relação aos meses anteriores. Caso o consumo diminua ou aumenta em relação a dois vezes a média dos últimos seis meses consecutivos, o consumidor é considerado suspeito de possível fraude.

$$\mathbf{VB = 2* \frac{x1+x2+x3\ldots x6}{N}}$$

Pode-se ainda classificar **consumidores como confiáveis e não confiáveis** junto a empresa baseado no pagamento da factura e o valor da conta (factura a ser paga).

Através destes indicadores podem ser seleccionados consumidores segundo regras a serem aplicadas sobre o volume de dados de investigação, combinando resultados individuais.

4.4.2. Planeamento Estratégico de obtenção, preparação e Análise dos dados

Definidos os indicadores, vamos descrever o processo de preparação da base de dados para a extracção do conhecimento. De realçar que a Base de Dados da Electra não está preparada para utilização de *Data Mining*, existe uma longa caminhada até que os dados transaccionais sejam transformados e armazenados (ETL) em uma *Data Warehouse* para interligar dados de forma centralizada disponibilizando-os para processos de extracção do conhecimento.

Para isso foram obtidas dados da base de dados dos consumidores fornecida pela Electra, relativa a algumas zonas da cidade da praia do ano de 2008. Foram analisadas quatro zonas geográficas com perfil dos clientes de diferentes classes sociais (duas de classe baixa, uma classe média e uma outra considerada de classe alta). Alguns dos factos que influenciaram o interesse por estas zonas para a descoberta do conhecimento reside com o objectivo de interrelacionar e comparar os dados dos clientes de diferentes classes sociais visando a identificação e descoberta de perdas de energia eléctrica bem como o perfil do cliente fraudulento.

Para validar a abordagem ao ambiente de dados das redes eléctricas, foi aplicado um algoritmo, árvore de decisão, simples e uma pequena amostra de dados dos clientes finais respectivas a facturação, consumo e anomalias.

Como referido anteriormente por (Fayyad, 1996), é necessário fazer a preparação dos dados de forma a evitar resultados erróneos. Para alcançar o melhor desempenho do algoritmo utilizado, foi necessário seleccionar os dados para definir quais características da rede eléctrica e do sistema de consumo e facturação eram pertinentes a análise pretendida e se possuíam alguma contribuição para a obtenção de resultados relevantes. Os dados foram extraídos da *Data Warehouse*, num total de 4800 registos, respeitantes á 400 clientes. Dos vários atributos utilizados, apenas 12 foram seleccionados pelos técnicos para análise como sendo interessantes.

Após certificar de que o conjunto era composto apenas por valores válidos, foi necessário prepara-los para oferecerem o máximo de significado e robustez na execução do algoritmo. Entre as transformações feitas, trabalhou-se sobre os atributos que calculavam as médias trimestrais e semestrais, as inspecções, estado e descrição das anomalias, os mínimos e máximos, etc.

Neste contexto foi utilizado o software *Weka,* versão 3.6.4 para análise dos dados provenientes das consultas no *Data Mart* e exportados para um ficheiro no formato".csv" separados por virgula. Depois convertido para um formato ".arff" *Attribute Relation Format File,* que é a extensão padrão utilizada pelo software *Weka* para a realização das tarefas de *Data Mining*, no qual tem que ser descrito o domínio dos atributos.

De referir que na fase de extracção e carregamento dos dados, os atributos e valores dos mesmos são carregados pela ferramenta Explorer, do software *Weka,* para que o algoritmo de classificação – arvore de decisão seja aplicado. Por fim, após a análise dos dados com

Data Mining, são interpretados os resultados do algoritmo aplicado.

Pode-se afirmar que a estrutura do arquivo "arff" é composta por relações, atributos e dados. A relação (*@relation*) é a primeira linha do ficheiro e deve conter a palavra reservada seguida de uma palavra-chave que identifique a relação ou a tarefa a ser analisada. Os atributos (*@attribute*) formam um conjunto de linhas onde cada uma contém a palavra reservada seguida do nome do atributo e do seu tipo que pode ser um *string* ou numérico. A última parte do arquivo "arff" corresponde ao conjunto dos valores dos dados (@data), inseridos logo após a definição dos atributos.

4.4.3. Interpretação e Análise dos Resultados

O algoritmo de *Data Mining* aplicado na classificação dos dados foi a Árvore de Decisão (J48) e tem como objectivo principal interpretar os dados previamente seleccionados a fim de produzir uma quantidade de padrões úteis, válidos e de fácil entendimento.

Os resultados gerados podem ser utilizados como predições e têm por finalidade a tomada de decisões inteligentes, visto que favorecem a extracção de informações de grandes volumes de dados e a análise estatísticas dos mesmos permitindo que se observem tendências e respostas para situações em que se possam detectar onde as perdas aparentes são mais frequentes, classificação dos clientes quanto a confiabilidade ou seja, quanto a irregularidades de pagamento das facturas e determinar o perfil dos consumidores quanto aos diferentes tipos de anomalias.

Essas regras foram aplicadas aos clientes de quatro tipos diferentes de zonas geográficas seleccionados previamente na análise dos postos de transformação em que duas dessas localidades foram consideradas pela empresa como sendo prioritária devido ao elevado índice de possíveis perdas de energia eléctrica. As outras zonas que foram analisadas são considerados pela empresa como sendo regulares mas poderá haver fraudes de energia eléctrica mesmo sabendo que o nível de vida é superior em comparação com as outras zonas.

4.4.4. Análise Prévia dos Resultados

Antes de iniciarmos o processo de *Data Mining* com o software *Weka*, o mesmo fornece algumas funcionalidades para fazer a realização da análise prévia dos dados e obtenção de informações importantes para o apoio a decisão, por meio de representações gráficas. Por exemplo pode-se fazer uma análise prévia dos tipos de anomalias existentes e a quantidade de valores encontradas para cada uma das anomalias numa das localidades mais problemáticas. Ao todo foram analisadas 1152 valores, sendo os tipos sem anomalias mais encontrados na base de dados (915).

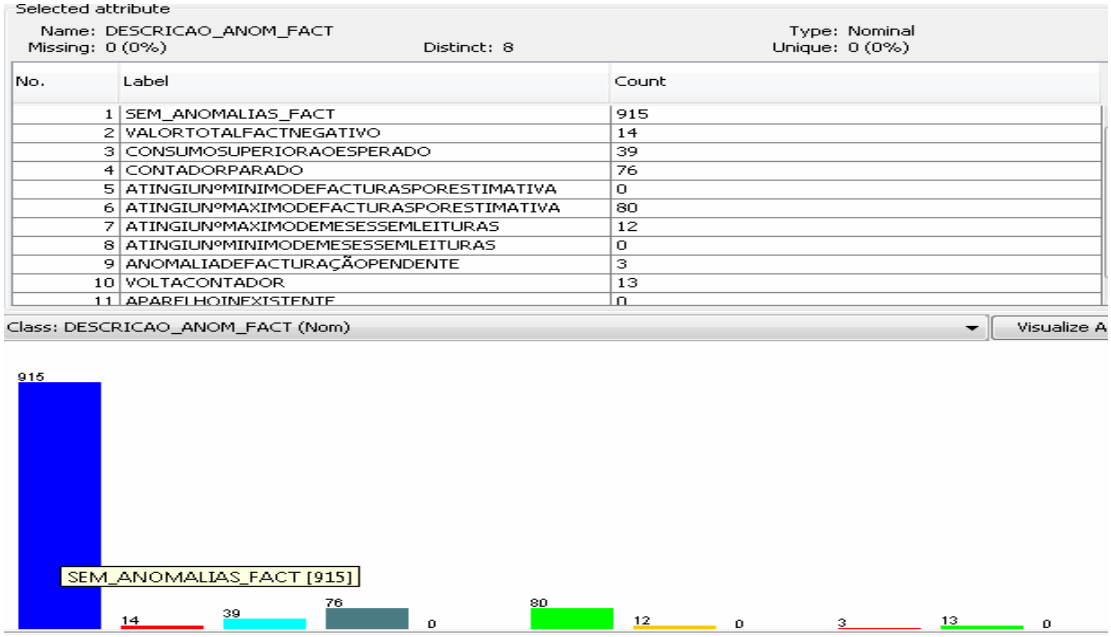


Gráfico 11 – Quantidade dos tipos de Anomalias

De acordo com o gráfico 12, permite-nos conhecer o perfil tarifário mais utilizados pelos clientes em que 936 valores são do tipo BTN (Baixa Tensão Normal), 204 BTE (Baixa Tensão Especial) geralmente utilizados pelas empresas e 12 considerados do tipo MT (Média Tensão), enquanto que o tipo de cliente ou categoria de consumo mais procurado são os domésticos com 1092 valores. Os clientes dessa zona consomem mais energia de baixa tensão normal com o tarifário pertencente a categoria de domésticos.

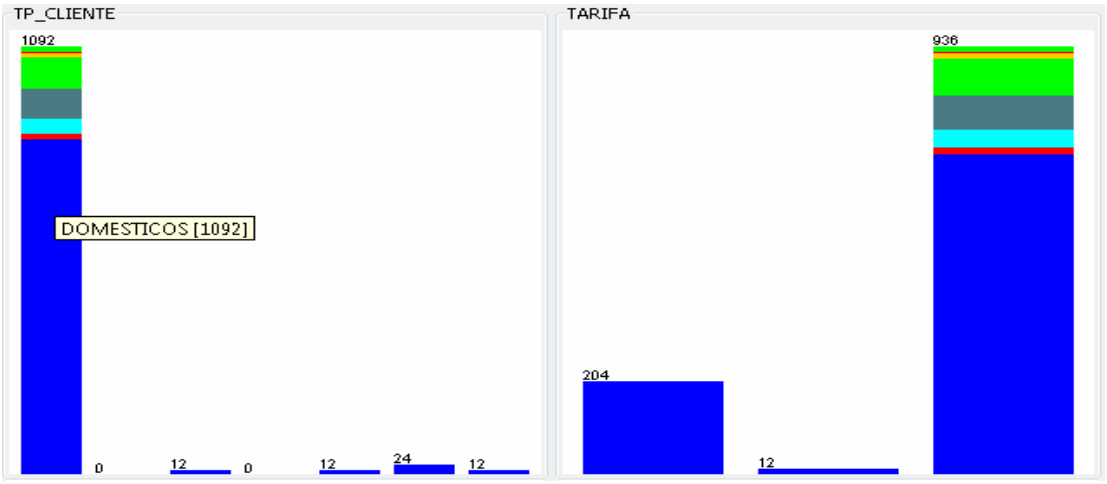
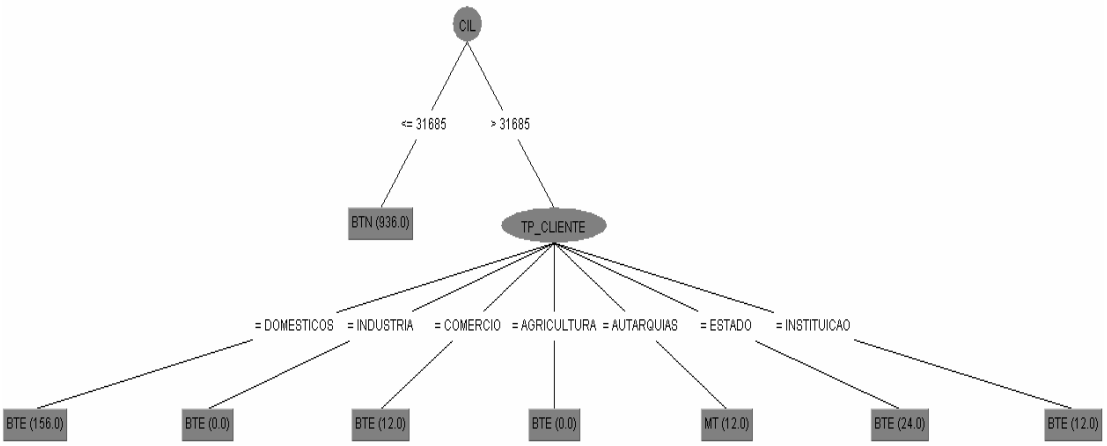


Gráfico 12 – Categoria do consumo e Tarifário mais utilizado

Um exemplo prático para obtermos conhecimento com as informações provindas destes gráficos, foi aplicado o algoritmo J48 em que a árvore de decisão do esquema 11, apresenta os tipos de clientes e o seu respectivo tarifário utilizado. Os clientes com o código de identificação local (CIL) menor ou igual a 31685 são do tipo baixa tensão normal (936), ou seja está identificada o tarifário com mais clientes desta localidade. Já os clientes com o CIL > 31685 pertencem a diferentes categorias de consumo e diferentes tarifários.



Esquema 11 – Categoria de consumo com o respectivo tarifário

4.4.5. Algoritmo Árvore de Decisão J48 aplicada na 1ª regra – Consumo Inferior ao Mínimo Esperado

A matriz de confusão faz parte da tarefa de classificação do *Data Mining*. A matriz dá uma informação muito útil porque reflecte os erros produzidos assim como informa sobre o tipo dos mesmos. Ela oferece meios efectivos para avaliação do modelo com base em cada classe cujo número de colunas é o número de atributos que mostra a classificação dos valores de cada atributo ou seja a classe predita, as linhas mostram a classe verdadeira enquanto que a diagonal principal representa os acertos e erros produzidos.

A tabela 3, ilustra um exemplo da matriz de confusão para um problema de uma classe classificada em (alto risco de fraudes, médio risco, baixo risco e situação regular).

=== CONFUSION MATRIX ===				
a	b	c	d	<-- classified as
219	0	0	0 |	a = ALTORISCO
0	195	0	0 |	b = MEDIORISCO
0	1	36	0 |	c = BAIXORISCO
1	0	0	700 |	d = REGULAR

Tabela 3 - Matriz de Confusão da 1ª regra

Pode-se realçar que ocorreram 219 casos de alto risco de fraudes, 195 casos de médio risco, 36 casos de baixo risco e 700 casos foram classificados como regulares e foram preditos como pertencentes à classe correcta.

A taxa de erro e a taxa de "*Accuracy"* são as medidas de avaliação mais utilizadas para os modelos de classificação visto que a taxa de erro é definida como sendo a proporção de erros de predição sobre um conjunto de exemplos em que se conhece o valor do atributo meta. Podemos dizer então que são estimativas do percentual de acertos e erros do classificador, respectivamente, na predição da classe de novos exemplos.

Neste contexto e de acordo com o exemplo abaixo respectiva a análise de uma zona problemática, podemos salientar que existem 1152 valores a serem classificadas de 97 clientes. Em que foram classificadas correctamente 1150 casos distribuídas pelas classes (alto, médio, baixo risco e regular) o que corresponde a uma taxa de *Accuracy* 99.8% dos casos. Foram dois, os valores classificados como incorrectos o que corresponde a uma taxa de erro de 0.18% dos casos verificados.

=== EVALUATION ON TRAINING SET === === SUMMARY ===		
Correctly Classified Instances	1150	99.8264 %
Incorrectly Classified Instances	2	0.1736 %
Total Number of Instances	1152	

Tabela 4 - Taxa percentual de erros e acertos de uma zona problemática

Utilizando a árvore de decisão J48 observa-se que a regra de classificação terá sempre no seu consequente uma resposta ao facto das condições satisfazerem ou não a uma determinada classe previamente definida.

Neste âmbito, os atributos dimensionais do *DataWarehouse* requeridos como entrada no algoritmo de *Data Mining*, árvore de decisão (J48) para geração e diagnóstico das perdas no consumo do cliente referentes a 1ª regra são (código de identificação do cliente, marca do contador, localidade, data do consumo, descrição da anomalias, media trimestral, situação do consumo, suspeitas de fraudes, inspecção). De referir que estes atributos e seus respectivos valores foram extraídos de diferentes *View´s* da *Data Mart* criada para este efeito através da tabela de factos e suas dimensões.

A figura seguinte ilustra a visão geral dos atributos da regra consumo abaixo do mínimo esperado que calcula as perdas no consumo de cada cliente de uma localidade problemática. Ao todo foram analisadas 96 clientes da referida localidade que corresponde a 1152 valores de 10 atributos.

De acordo com a figura 13, todos os gráficos estão associados ao atributo de decisão "Inspeccionar" sendo a cor azul corresponde ao tipo "Não Inspeccionar" e a cor vermelha corresponde ao tipo "Inspeccionar". Dos 1152 valores dos atributos 452 são classificadas para fazer a inspecção na referida zona e 700 estão com a sua situação regular não inspeccionar.

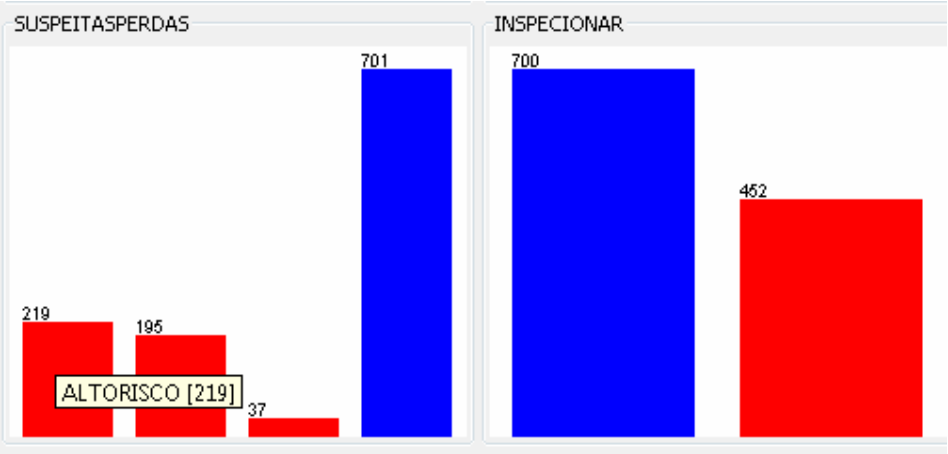


Gráfico 13 – Classificação dos clientes

O atributo inspeccionar foi utilizado como atributo classe (atributo padrão) e desta forma todos os atributos do modelo estabelecem relação directa com este atributo. O gráfico da figura 13 "Suspeitas de Perdas" apresenta 219 casos de alto risco com elevada urgência de prioridade de selecção para inspecção, 195 casos de médio risco, 37 casos de baixo risco todos destacados a vermelho e 701 casos regulares destacado a azul.

De salientar que um cliente pode ser suspeito de perdas de alto risco (prioridade elevada para inspecção) se tiver um consumo abaixo do mínimo esperado e sem anomalias. É considerado de médio risco se tiver um mínimo esperado com anomalias. De baixo risco se não tiver um mínimo esperado ou seja situação normal e tiver anomalias de facturação e leituras. É regular se apresenta com a situação do consumo normal e sem anomalias.

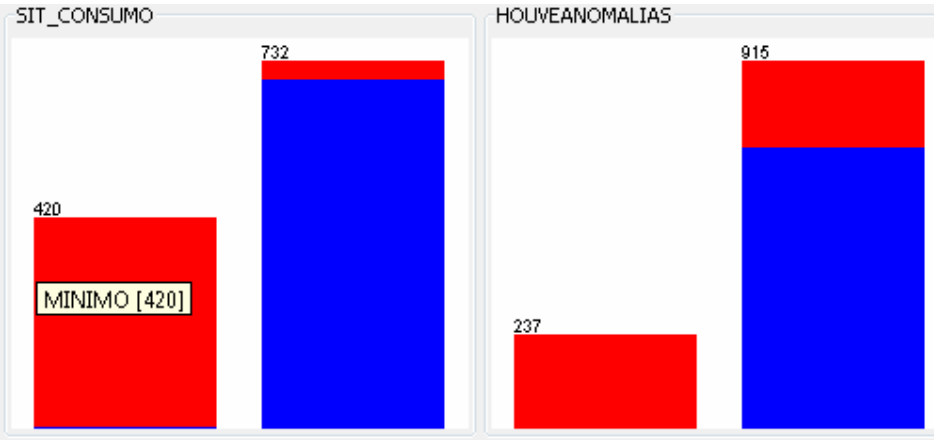
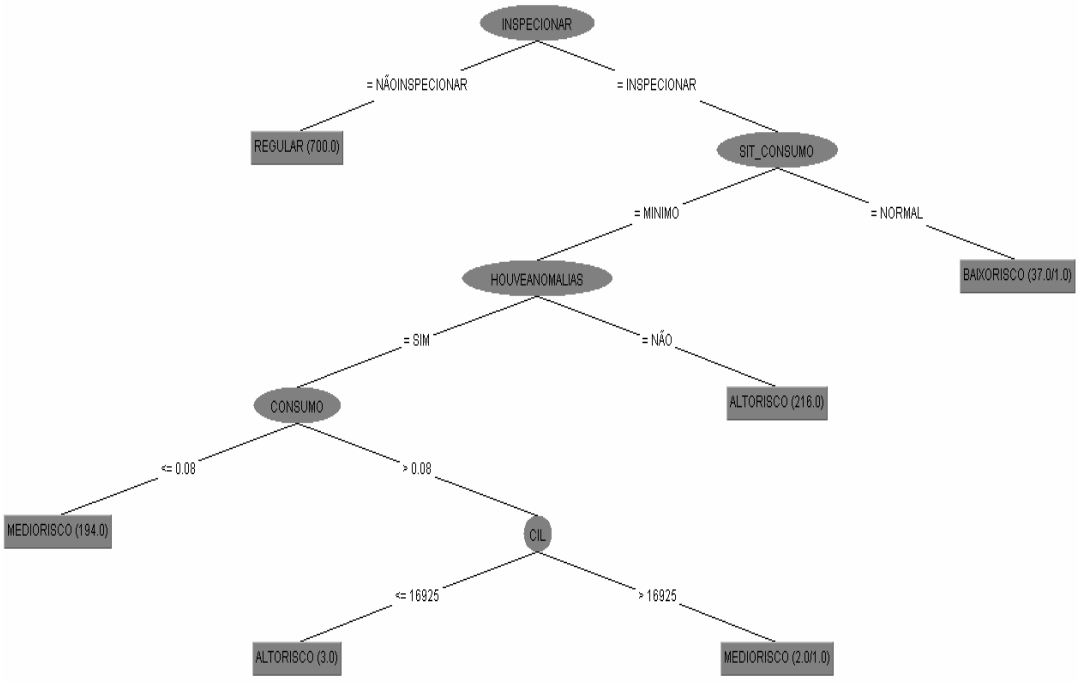


Gráfico 14 – Consumo abaixo do mínimo esperado e número de anomalias

A partir deste gráfico, pode-se observar que houve 420 casos de clientes com vários meses com o consumo abaixo do mínimo esperado, são considerados suspeita de possíveis fraudes de energia eléctrica. Calcula-se o consumo mínimo esperado como sendo inferior á 0,6 (valor atribuído pelos técnicos) vezes média dos últimos três meses. Caso tenha um mínimo esperado verifica se houve anomalias de leitura e facturação nesse período, se sim, classifica-as de baixo, médio e alto risco de fraudes para inspecção por ordem de prioridade, caso não tenha anomalias, é seleccionado imediatamente para inspecção.



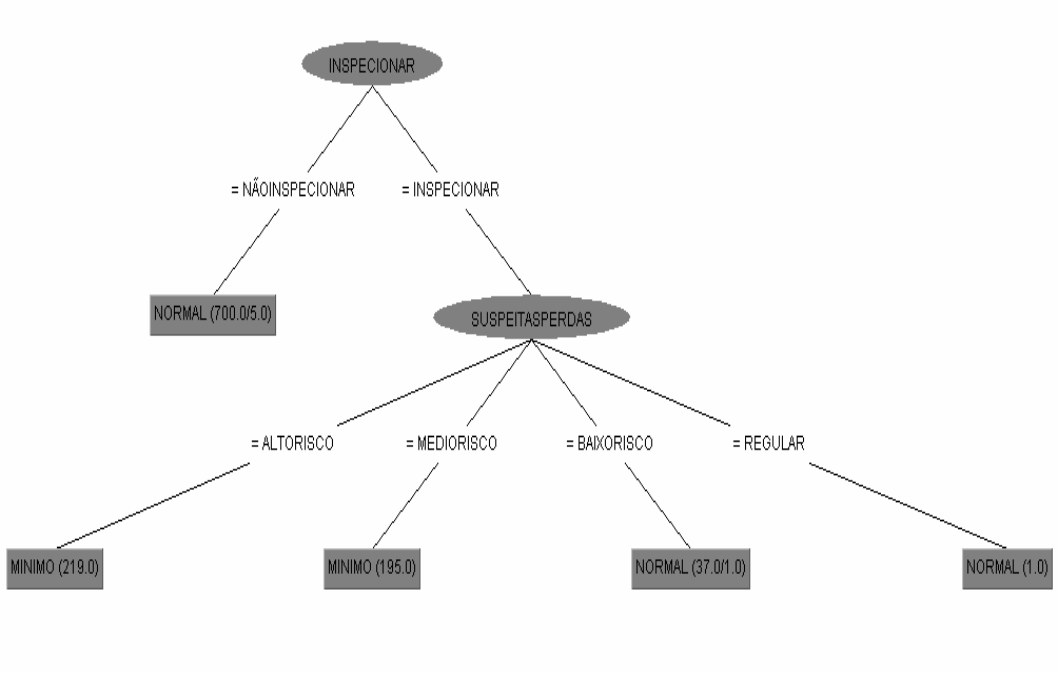
Esquema 12 – Clientes seleccionados para inspecção

Um cliente é seleccionado para inspecção se tiver um consumo inferior ao mínimo esperado e não houver anomalias, é classificada como sendo de alto risco. Foram encontradas 216 casos de alto risco. Se houver anomalias verifica-se se o consumo atinge o mínimo esperado e classifica-os como sendo de alto e médio risco. Foram encontradas 194 casos de médio risco. Se a situação for normal é classificado como sendo de baixo risco (36 casos) e um erro, deveria ser classificado como sendo de médio risco. Não inspeccionar se a situação for regular (700 casos) e um erro, deveria ser classificado como sendo de baixo risco.

```
J48 PRUNED TREE
    INSPECIONAR = NÃOINSPECIONAR: REGULAR (700.0)
    INSPECIONAR = INSPECIONAR
    |  SIT_CONSUMO = MINIMO
    |  |  HOUVEANOMALIAS = SIM
    |  |  |  CONSUMO <= 0.08: MEDIORISCO (194.0)
    |  |  |  CONSUMO > 0.08
    |  |  |  |  CIL <= 16925: ALTORISCO (3.0)
    |  |  |  |  CIL > 16925: MEDIORISCO (2.0/1.0)
    |  |  HOUVEANOMALIAS = NÃO: ALTORISCO (216.0)
    |  SIT_CONSUMO = NORMAL: BAIXORISCO (37.0/1.0)
```

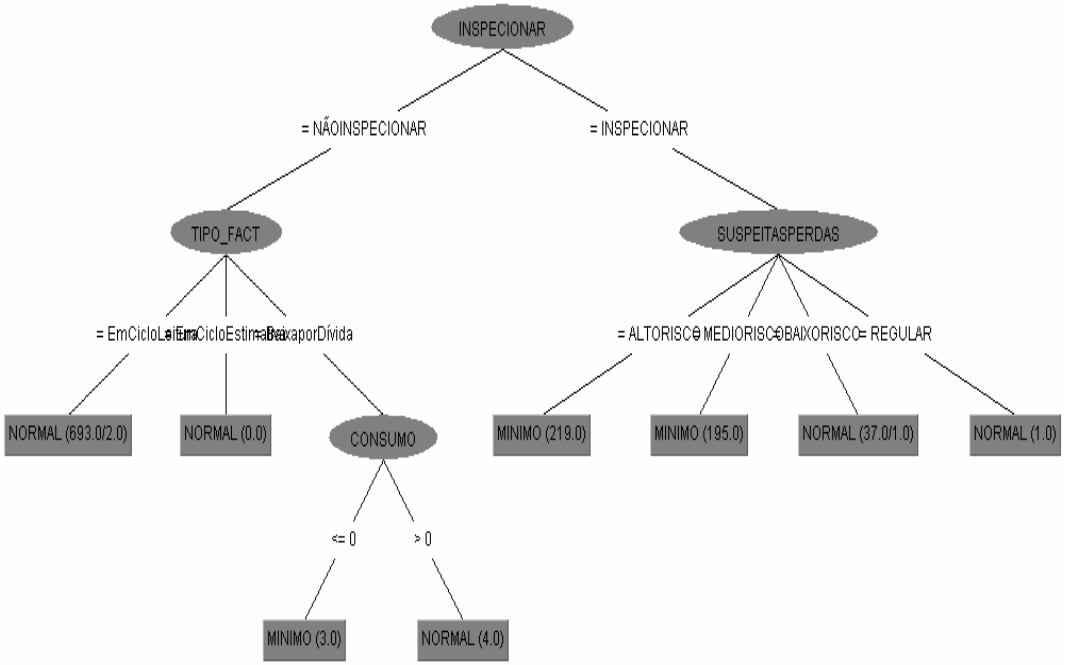
Tabela 5 – Árvore de Decisão

Partindo-se do nó raiz obtemos como padrão de comportamento para este primeiro caso que todos os clientes com a situação de consumo de energia regular não são inspeccionados. Quando são seleccionados para inspecção outros nós (atributos) precisam ser verificados. Por exemplo, todos os clientes que obtiveram um consumo abaixo do mínimo esperado e tiveram anomalias são classificados para inspecção por ordem de prioridade de acordo com a gravidade dos riscos de perdas. Se obtiveram um mínimo esperado sem anomalias são imediatamente classificados para inspecção com elevado grau de encontrar perdas de energia eléctrica.



Esquema 13 - Mínimo Esperado

A interpretação de todas as árvores de decisão determina todos os padrões descobertos da base de dados. A árvore do esquema 14, determina que os clientes para serem inspeccionados, devem ser suspeitos de fraudes por possuírem um consumo abaixo do mínimo esperado. Caso se encontrem numa situação normal verifica-se se a facturação está em ciclo de leitura (normal), em ciclo estimativa (por ausência ou impossibilidade de leituras), baixa por dívida (cortes por não pagamento da factura) ou consumo inferior a zero.



Esquema 14 - Arvore de decisão mínimo esperado por tipo de facturação

Os resultados dessa experiência são considerados positivos pois mantiveram-se com confiabilidade negativa superior aos índices que os técnicos possuem nas actuais formas de selecção dos clientes para inspecção nessa referida zona, muito embora faltam ainda os testes de validação da regra no terreno. Com estes padrões identificados pode-se seleccionar clientes para inspecção por ordem de prioridade de risco de perdas eléctricas.

4.4.6. Algoritmo Árvore de Decisão J48 aplicado na 2ª regra – Confiança dos Clientes

Para classificar os clientes como confiáveis e não confiáveis junto a empresa baseado no pagamento das facturas e valor das contas foram utilizados 10 atributos: o código de identificação do cliente, tipo cliente, descrição e estados das anomalias, tipo de facturação, tarifa utilizada, estado da facturação bem como a classificação quanto a confiabilidade dos clientes.

Ao executar o algoritmo J48 com os dados da confiabilidade dos clientes, o software *Weka* gerou as informações abaixo descritas de forma sucinta. Ao todo foram utilizadas 1152 valores de 10 atributos para a classificação, sendo 1149 valores, classificadas correctamente (taxa de *Accuracy* 99,7%), 3 valores classificadas, incorrectamente (taxa de erro 0.26%). De referir que a taxa de A*ccuracy* 99,7% indica a alta precisão na classificação dos dados, reflectindo resultados confiáveis sobre os dados analisados.

```
=== RUN INFORMATION ===
------------------
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     CONFIABILIDADE
Instances:    1152
Attributes:   10
Correctly Classified Instances        1149               99.7396 %
Incorrectly Classified Instances         3                0.2604 %

=== Confusion Matrix ===

   a   b   c   <-- classified as
 885   0   0 |   a = COBRADA
   2 241   0 |   b = NCOBRADA
   1   0  23 |   c = PORCOBRAR
```

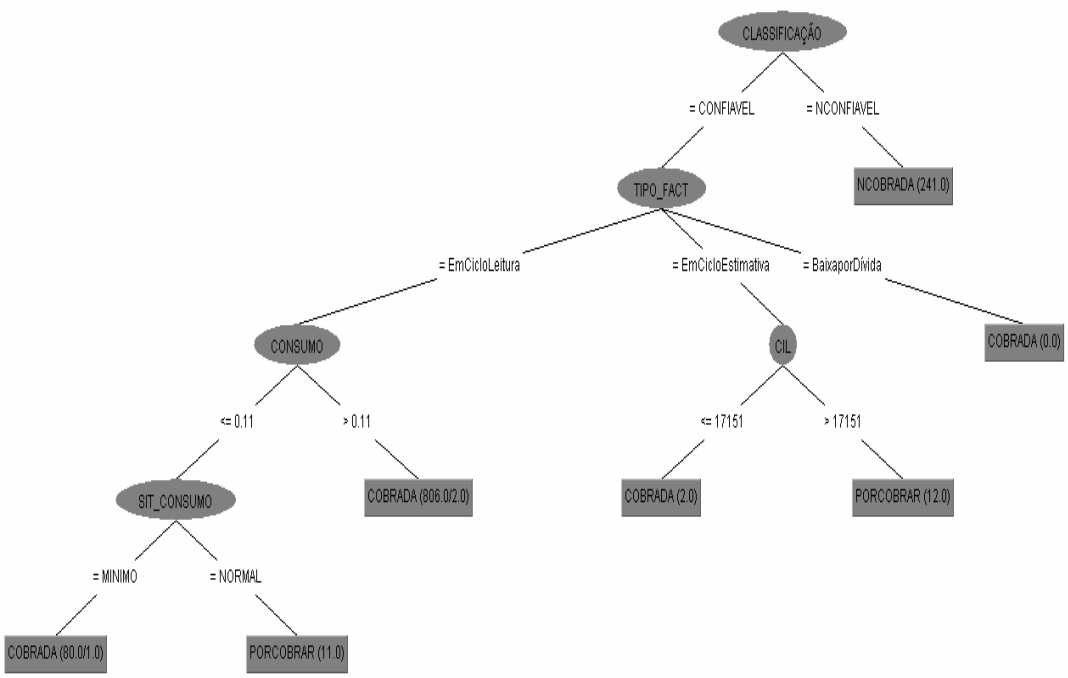
Tabela 6 – Matriz de Confusão da 2ª regra

A matriz de confusão produzida pelo algoritmo J48 mostra que dos 1152 valores classificados, 1149 foram classificados correctamente e 3 classificados incorrectamente. Ou seja, 885 casos foram classificados como cobrados, portanto confiáveis, 241 casos não cobrados e 2 omissões, visto que deveriam ter sido classificados como cobrados e 23 casos por cobrar (em ciclo estimativa ou por outro motivo).

```
J48 PRUNED TREE
    ------------------
    ------------------
    CLASSIFICAÇÃO = CONFIAVEL
    |  TIPO_FACT = EmCicloLeitura
    |  |  CONSUMO <= 0.11
    |  |  |  SIT_CONSUMO = MINIMO: COBRADA (80.0/1.0)
    |  |  |  SIT_CONSUMO = NORMAL: PORCOBRAR (11.0)
    |  |  CONSUMO > 0.11: COBRADA (806.0/2.0)
    |  TIPO_FACT = EmCicloEstimativa
    |  |  CIL <= 17151: COBRADA (2.0)
    |  |  CIL > 17151: PORCOBRAR (12.0)
    |  TIPO_FACT = BaixaporDívida: COBRADA (0.0)
    CLASSIFICAÇÃO = NCONFIAVEL: NCOBRADA (241.0)
```

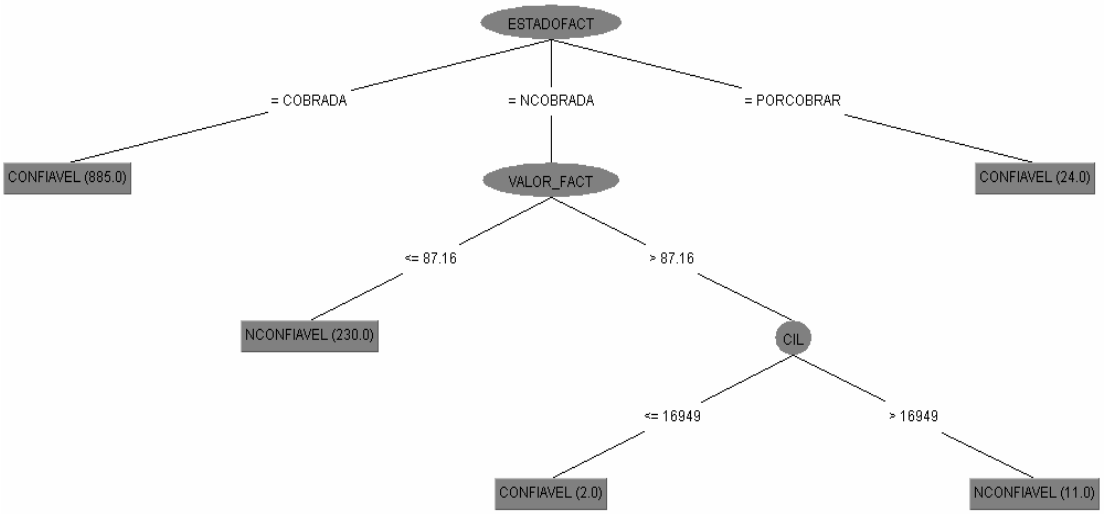
Tabela 7– Arvore de decisão da 2ª regra

As regras de classificação correspondem ao conhecimento descoberto e são geradas em forma de árvore de decisão, conforme mostra o esquema 15. Desta forma todos os clientes que têm o estado de factura não cobradas nos últimos três meses são classificados como não confiáveis. De realçar que os clientes que estão associadas a esta regra podem estar a causar perdas de energia eléctrica e a solução proposta é a inspecção desses mesmos clientes. Um aspecto interessante é que foram encontradas 80 casos classificados como confiáveis, em ciclo de leitura, com um consumo abaixo do mínimo esperado e cobradas. Este caso foi avaliado pelos técnicos da empresa como sendo clientes com vários meses ausentes ou sem consumo por uma outra razão e são cobradas as taxas mínimas.



Esquema 15 – Clientes Confiáveis e não Confiáveis

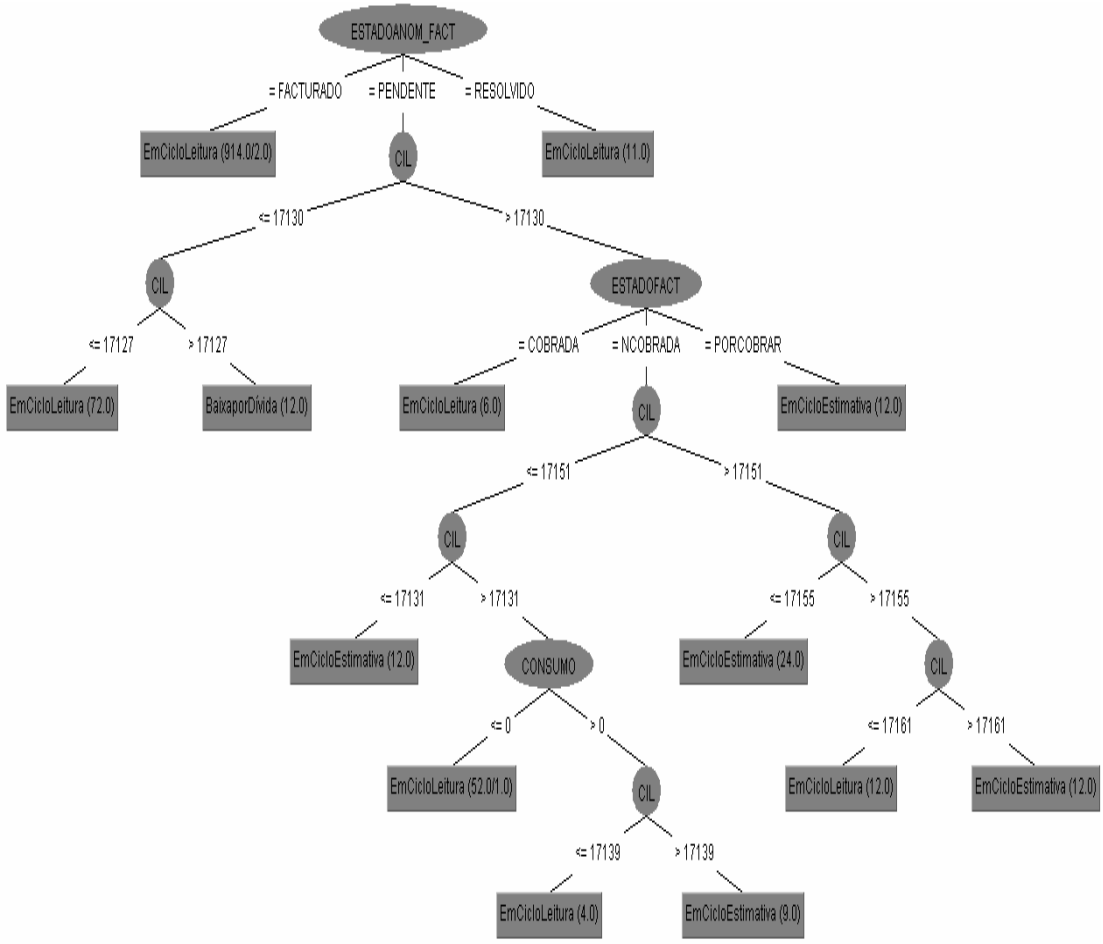
Outro padrão encontrado nesta mesma localidade foi que os clientes com o estado de factura não cobrada e com o valor da factura menor que 87.16 são considerados não confiáveis. Foram encontradas 885 casos de confiança quando a situação é cobrada, 24 por cobrar e 230 casos não cobradas quando o valor da factura é menor ou igual a 87.16 escudos portanto classificados como não confiáveis. De referir que foram encontradas também dois casos de confiança quando não pagaram cujo valor da factura é maior que 87.16 e 11 casos não confiáveis.



Esquema 16 – Estado da Facturação

A figura 17 apresenta uma árvore de decisão em que foram relacionados o estado das anomalias de facturação e o estado de facturação das facturas dos clientes. Verifica-se que 1068 registos foram classificados em ciclo de leituras (normal), 69 casos em ciclo estimativa e três registos classificadas incorrectamente, visto que deveriam ser classificadas como em ciclo de leituras e 12 casos de baixa por dívida.

Como padrões identificados nesta árvore, verificam-se casos em que um cliente é facturado quando está em ciclo de leitura, se o estado é pendente verifica se a factura foi cobrada, não cobrada ou está por cobrar. Não cobrada e por cobrar podem estar em ciclo estimativa ou baixa por dívida.



Esquema 17– Relação entre estado das anomalias e das facturas

Os resultados com a aplicação desta regra pode conduzir os técnicos a descobrir padrões de confiabilidade dos clientes em relação ao pagamento das facturas. De salientar que o não pagamento das facturas pode levar o cliente a praticar fraudes.

4.4.7. Algoritmo Árvore de Decisão J48 aplicado na 3ª regra – Variação brusca do consumo

Ao executar o algoritmo com os dados da regra variação brusca, o *Weka* gerou os resultados apresentados na tabela abaixo. No total foram utilizados 1152 amostras de uma localidade problemática e 11 atributos para a classificação. Sendo 1065 valores classificadas correctamente (taxa *Accuracy* 92.4479%) e 87 valores classificadas incorrectamente (taxa de erro 7.5521). De realçar que o tempo de processamento foi de 0,16 segundos.

```
=== SUMMARY ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     INSPECIONAR
Instances:    1152
Attributes:   11
Time taken to build model: 0.16 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1065               92.4479 %
Incorrectly Classified Instances        87                7.5521 %
=== Confusion Matrix ===

   a   b   c   <-- classified as
 416   0   8 |   a = MINIMO
   0  41  74 |   b = MAXIMO
   2   3 608 |   c = NORMAL
```

Tabela 8 - Matriz de confusão da 3ª regra

A matriz de confusão produzida pelo algoritmo classificador J48 mostra que 416 registos foram classificados correctamente com uma variação brusca para o mínimo e 8 como incorrectas. Quanto ao tipo de variação brusca com tendência para o máximo foram classificadas 41 registos como sendo correctas e 74 registos foram classificadas incorrectamente visto deveriam ser classificadas como sendo normal. Para os casos normais foram classificadas 608 valores correctamente e 5 incorrectas.

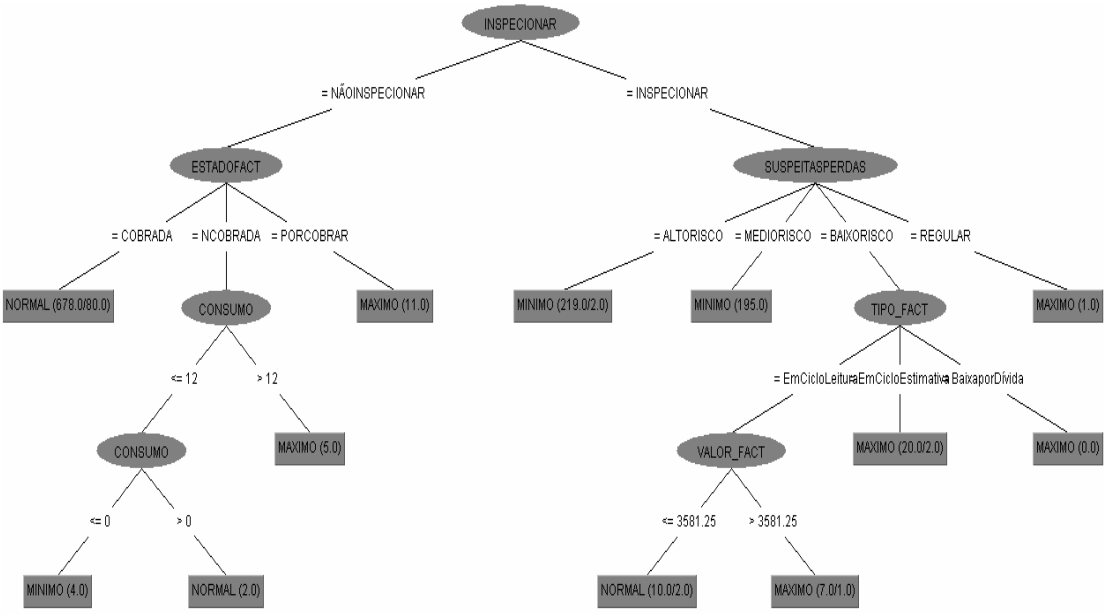
```
J48 PRUNED TREE
INSPECIONAR = NÃOINSPECIONAR
|  ESTADOFACT = COBRADA: NORMAL (678.0/80.0)
|  ESTADOFACT = NCOBRADA
|  |  CONSUMO <= 12
|  |  |  CONSUMO <= 0: MINIMO (4.0)
|  |  |  CONSUMO > 0: NORMAL (2.0)
|  |  CONSUMO > 12: MAXIMO (5.0)
|  ESTADOFACT = PORCOBRAR: MAXIMO (11.0)
INSPECIONAR = INSPECIONAR
|  SUSPEITASPERDAS = ALTORISCO: MINIMO (219.0/2.0)
|  SUSPEITASPERDAS = MEDIORISCO: MINIMO (195.0)
|  SUSPEITASPERDAS = BAIXORISCO
|  |  TIPO_FACT = EmCicloLeitura
|  |  |  VALOR_FACT <= 3581.25: NORMAL (10.0/2.0)
|  |  |  VALOR_FACT > 3581.25: MAXIMO (7.0/1.0)
|  |  TIPO_FACT = EmCicloEstimativa: MAXIMO (20.0/2.0)
|  |  TIPO_FACT = BaixaporDívida: MAXIMO (0.0)
|  SUSPEITASPERDAS = REGULAR: MAXIMO (1.0)
```

Tabela 9 - Arvore de decisão da 3ª regra

Ao analisar a árvore de decisão gerada pela execução do algoritmo J48 é possível descobrir padrões de classificação. Para seleccionar um cliente para inspecção, ele deverá ser suspeito de fraudes e ter uma variação brusca de consumo em dois meses consecutivos. Conforme indica a árvore se o nó raiz inspeccionar for = inspeccionar e suspeitas perdas for de alto e médio risco com a folha mínimo esperado. Esta regra informa que todos os clientes que tiveram um consumo inferior ao mínimo esperado são suspeitos de perdas de alto risco com alta e média prioridade de inspecção.

Um outro padrão descoberto na árvore, nó raiz inspeccionar = inspeccionar, suspeitas perdas = baixo risco, tipo facturação = em ciclo de leitura e o valor da factura> 3581.25. Esta regra mostra que um cliente pode ser inspeccionado se a suspeita de fraude é baixa, está em ciclo de leitura e se o seu valor da factura é maior que o intervalo de valor dado para uma determinada tarifa.

De salientar também que caso o cliente não seja seleccionado para inspecção devido a se encontrar numa situação normal, é verificado o estado da factura (cobrada, não cobrada, por cobrar). Se for não cobrada é verificado o seu consumo, e se este for <= 0 o cliente pode estar ausente ou não.



Esquema 18 – Variação Brusca do Consumo

4.4.8. Discussão e Avaliação dos resultados

O pressuposto inicial de que há uma grande quantidade de informação e conhecimento ocultos nos registos da base de dados da Electra, é válido, uma vez que a riqueza das informações obtidas a partir das respostas alcançadas ao longo deste trabalho são um bom suporte para a tomada de decisão perante as perdas.

Respondendo à pergunta de partida elaborada no inicio do estudo de caso "Quais os padrões de informação existentes na Base de Dados, que permitem a descoberta de indícios, evidências ou, pelo menos, suspeitas da ocorrência de perdas de energia eléctrica" foi possível encontrar e confirmar diversos padrões de comportamento.

Por exemplo, na regra **consumo inferior ao mínimo esperado**, um cliente classificado como alta prioridade de inspecção por perdas aparentes deve apresentar um consumo inferior a um mínimo esperado e sem anomalias de facturação ou de leituras. Neste âmbito o cliente é seleccionado como tendo um perfil de alta prioridade de inspecção, suspeita de perdas.

Na mesma regra, se o consumo for inferior ao mínimo esperado e existirem anomalias de facturação ou de leituras, o cliente é classificado como sendo de médio risco de perdas de energia eléctrica. Finalmente, um cliente é considerado de baixo risco se tiver um

consumo dentro do esperado, mas existirem anomalias de facturação ou de leituras. De realçar que estas anomalias podem ser causadas por infracções no contador e assim alterar o perfil do cliente. É regular se o cliente apresenta com a situação do consumo normal e sem anomalias.

Para a segunda regra, **classificação dos clientes como confiáveis e não confiáveis** junto a empresa baseado no pagamento das facturas e valor das contas, foram identificadas também alguns padrões que poderão nos levar a perdas de energia eléctrica.

Desta forma todos os clientes que têm o estado de factura não cobrada nos últimos três meses são classificados como não confiáveis. De realçar que os clientes que estão associadas a esta regra podem estar a causar perdas de energia eléctrica e a solução proposta é a inspecção desses mesmos clientes.

Outro padrão encontrado nesta mesma localidade foi que os clientes com o estado de factura não cobrada e com o valor da factura menor que 87.16 são considerados não confiáveis.

Já na terceira regra, **variação brusca do consumo**, o perfil dos clientes a seleccionar para inspecção por serem suspeitos de fraudes, devem apresentar uma variação brusca de consumo, tendendo para o máximo ou para o mínimo, em dois meses consecutivos. Se a variação tender para o mínimo há alto risco de perdas de energia eléctrica; se tender para o máximo e existirem anomalias, o cliente é considerado de médio risco de fraudes, senão, é considerado de baixo risco de fraudes.

Um outro padrão descoberto na árvore, mostra que um cliente pode ser inspeccionado se a suspeita de perdas é baixa, está em ciclo de leitura e se o seu valor da factura é maior que o valor máximo do intervalo estimado para uma determinada tarifa.

De salientar que para os clientes que não são seleccionados para inspecção por serem considerados numa situação normal, é verificado o estado da factura (cobrada, não cobrada, por cobrar). Se o estado for não cobrado, verifica-se se o seu consumo é <= 0 considera-se a possibilidade de o cliente estar ausente, se o estado for por cobrar, verifica-se se tem alguma anomalia, caso contrario é cobrado normalmente.

De referir que para os casos das zonas consideradas pelos técnicos como mais problemáticas e de classe social baixa, o índice das perdas foi semelhante aos índices que os técnicos obtiveram usando as formas actuais de selecção e realização de inspecções. No entanto, falte ainda analisar esses mesmos dados no terreno ou seja a validação dos

mesmos.

Foram utilizados procedimentos adequados ao processo de descoberta do conhecimento, e destas experiências obtivemos resultados satisfatórios. Acreditamos pelas razões já citadas, que os resultados tendem a melhorar, na medida em que podem ser utilizados para novas experiências um volume de dados maior e inserção de mais regras. Mesmo, assim os resultados apresentados demonstram uma melhora significativa frente aos métodos tradicionais utilizados para a identificação de perdas.

É bom lembrar que actualmente, o combate as perdas de energia eléctrica é feita através de inspecções em zonas ou localidades geográficas sem terem uma directriz de busca, essas inspecções no terreno nem sempre conseguem obter o flagrante de todas as perdas devido ao tempo de demora, alto custo envolvido e devidas as leis vigentes. Com este sistema de apoio a decisão os técnicos têm o conhecimento dos postos de transformação e as áreas criticas mais problemáticas e selecciona os seus clientes finais para serem inspeccionados.

Em termos quantitativos, os índices de acertos em inspecções no terreno, ou seja pelo método tradicional, estão entre 10 e 25%, dependendo da região e do tamanho das equipas de serviço mobilizadas. O nosso estudo permitiu-nos obter, resultados na ordem dos 30 a 47%.

Foram analisadas os dados de zonas geográficas de diferentes classes sociais, abaixo descritas na tabela 10, relacionando e comparando os resultados finais obtidos para cada uma das regras aplicadas nas referidas zonas.

Pode-se observar na tabela que a Zona de Eugénio Lima com 1152 registos analisados respeitantes a 97 clientes, apresenta com um grau elevado de inspecções nas residências cerca de 39%, em que 19% são consideradas de elevada prioridade de inspecção. Mas no entanto em relação à facturação, 79% dos clientes são consideradas confiáveis pelo pagamento das suas facturas. De salientar que um cliente pode ser assíduo no pagamento das suas facturas e ser um cliente fraudulento. Em relação à variação brusca do consumo obtivemos cerca de 36% de casos a tender para o valor mínimo o que demonstra a necessidade premente de inspecção nesta localidade.

Uma outra zona problemática analisada (Pensamento), mostra também um elevado grau de suspeitas de perdas, referentes a 47% dos 1284 registos analisados respectivos a 107 clientes. Este aumento de perdas em relação à zona de Eugénio Lima deve-se também a uma quantidade maior de casos analisados.

Constata-se ainda que, os percentuais de perdas ou de selecção para inspecção e o índice de confiança diminuíram ligeiramente em relação à zona de classe alta e significativamente em relação à zona de classe média. Um facto notado com a análise, é que a zona do Palmarejo classificada como sendo uma zona de classe alta teve um percentual de perdas e de confiabilidade maior que a zona da Terra Branca classificada de classe média.

ZONAS GEOGRAFICAS	**REGRAS**	**SITUAÇÃO DO CONSUMO**		**PERDAS APARENTES POR ORDEM DE PRIORIDADE DE INSPECÇÃO**		
Zona Problemática classe baixa (Eugénio Lima)	**Mínimo Esperado**	Normal		Alta	Média	Baixa
		700		219	195	36
	Confiança	Confiável	Não Confiável	Cobradas	Não cobradas	Por cobrar
		911	241	885	241	23
	Variação Brusca do Consumo	Normal		Mínimo	Máximo	
		608		416	41	

Zona problemática classe baixa (Pensamento)	**Mínimo Esperado**	Normal		Alta	Média	Baixa
		616		160	244	256
	Confiança	Confiável	Não Confiável	Cobradas	Não cobradas	Por cobrar
		1033	251	936	250	96
	Variação Brusca do Consumo	Inspeccionar	Não Inspeccionar	Normal	Mínimo	Máximo
		627	651	695	331	215

Zona Palmarejo (Classe Alta)	**Mínimo Esperado**	Normal		Alta	Média	Baixa
		753		193	116	130
	Confiança	Confiável	Não Confiável	Cobradas	Não cobradas	Por cobrar
		1029	170	993	121	84
	Variação Brusca do Consumo	Inspeccionar	Não Inspeccionar	Normal	Mínimo	Máximo
		537	649	308	146	721

Zona Terra Branca (Classe Média)	**Mínimo Esperado**	852		Alta	Média	Baixa
				192	59	90
	Confiança	Confiável	Não Confiável	Cobradas	Não Cobradas	Por Cobrar
		1089	111	1005	111	84
	Variação Brusca do Consumo	Inspeccionar	Não Inspeccionar	Normal	Mínimo	Máximo
		431	754	800	253	130

Tabela 10 - Dados estatísticos das regras aplicadas nas zonas

A taxa de acertos e de erros do algoritmo aplicado para a primeira regra nas diferentes zonas tiveram uma média de 99,46% e 0,49% respectivamente. Para a segunda regra a média da taxa de acertos foi de 99,85% e taxa de erro de 0.12%. A terceira regra apresenta um total médio de acerto de 96,5% e de erro de 3,44. Algumas regras aplicadas tiveram taxas de acertos e erros superiores em zonas diferentes devido ao número de atributos e

valores.

ZONAS GEOGRAFICAS	**REGRAS**	**TAXA DE ACERTOS**	**TAXA DE ERROS**	
Zona Problemática classe baixa (Eugénio Lima)	Mínimo Esperado	99,8	0,18	1152 Registos, 11 atributos
	Confiança	99,7%	0,26%	1152 Registos, 11 atributos
	Variação Brusca do Consumo	92,44%	7,5%	1152 Registos, 11 atributos
Zona problemática classe baixa (Pensamento)	Mínimo Esperado	99.37%	0.62%	1284 Registos, 11 atributos
	Confiança	99.9%	0.07%	1284 Registos, 10 atributos
	Variação Brusca do Consumo	96.6%	3.3%	1284 Registos, 10 atributos
Zona Palmarejo (Classe Alta)	Mínimo Esperado	99,3%	0,66%	1200 Registos, 11 atributos
	Confiança	99,8%	0.16%	1200 Registos, 10 atributos
	Variação Brusca do Consumo	98,41%	1,58%	1200 Registos, 10 atributos
Zona Terra Branca (Classe Média)	Mínimo Esperado	99,4	0,5	1200 Registos, 11 Atributos
	Confiança	100%	0%	1200 Registos, 10 atributos

	Variação Brusca do Consumo	98,58	1,41	1200 Registos, 10 atributos

Tabela 11 - Taxa de acertos e Erros das zonas seleccionadas para inspecção

Após a avaliação dos resultados podemos frisar que a aplicação do algoritmo de *Data Mining* J48, visando os objectivos propostos anteriormente foi alcançada, proporcionando entre outras funcionalidades e benefícios, a análise e descoberta de padrões que nos leva a detecção de perdas de energia nas localidades seleccionadas, detecção de anomalias relacionando-as ao perfil do cliente.

Os resultados alcançados precisam ser validados pelos técnicos da empresa com inspecções no terreno. Após a validação por parte dos técnicos, o estudo poderá ser colocado em prática pela empresa adaptando o mesmo para o efeito, visto que necessita de se trabalhar com uma equipa interdisciplinar, relacionando os conhecimentos dos técnicos com aqueles que dominam as tecnologias, em especial *Data Mining*.

5. Conclusões e Recomendações

Recentemente tem acontecido um aumento crescente de perdas de energia eléctrica em diversos sectores direccionados a um grande público e, em particular, na empresa de produção e distribuição de energia eléctrica (Electra). Este tipo de irregularidade representa uma forma de subtrair ilicitamente bens e serviços alheios em benefício próprio, afectando o orçamento do provedor desses serviços. Por isso, actualmente, a empresa de produção e distribuição de energia eléctrica busca soluções que auxiliem a detecção de fraudes de energia eléctrica.

Neste contexto uma das principais motivações para realização deste trabalho foi o de investigar detalhadamente os dados da base de dados que compõe o sistema de distribuição de energia eléctrica com o intuito de identificação de perdas de energia eléctrica nos postos de transformação e clientes finais, utilizando as tecnologias de *Data Warehouse* para selecção, organização e armazenamento dos dados extraídos da base de dados da Electra, para em seguida serem aplicadas as tecnologias *OLAP* e a técnica de *Data Mining* para extracção de padrões de conhecimento.

Este trabalho proporcionou uma visão das características reais da rede eléctrica da Cidade da Praia, de como estão organizados os dados das redes de distribuição de energia, dos postos de transformação, bem como os dados da facturação e consumo dos clientes, e permitiu identificar deficiências existentes, e o que pode ser mudado ou melhorado com a implementação de um Sistema de Apoio a Decisão.

O desenvolvimento deste estudo possibilitou-nos adquirir experiência e fazer uma

análise mais abrangente sobre os dados extraídos da base de dados para um *Data Warehouse*, que por sua vez, oferece os fundamentos e os recursos necessários para um sistema de suporte a decisão eficiente, fornecendo dados integrados e históricos. Este mostrou-se de grande valia quando aplicado ao sistema, principalmente no processo de integração e gestão dos dados extraídos de diversas fontes, com o propósito de visualizar os dados em diferentes níveis de detalhe e classificá-los usando critérios de selecção dos dados adequados ao caso de estudo.

De referir que a ideia inicial era utilizar todos os dados da base de dados dos clientes de algumas zonas facultados pela empresa, mas verificou-se que o processo de preparação, tratamento e transformação dos dados seria muito lento, consumindo a maior parte do tempo que poderia ser dedicado ao processo de descoberta do conhecimento, assim optou-se por utilizar como amostra de dados cerca de 400 clientes (domésticos e empresas) o que equivale a 4800 linhas de dados ou registos referentes a quatro zonas de diferentes classe sociais (duas de classe baixa, uma de classe média e outra de classe alta) respeitantes ao ano de 2008.

Foi possível, ainda, perceber de forma clara as vantagens da utilização do *Data Warehouse* e *OLAP*, para organização, integração e visões personalizadas dos dados dos transformadores e clientes finais. Estas vantagens das tecnologias *Data Warehouse* e *OLAP* permitiu-nos identificar previamente os postos de transformação e as áreas criticas com maior incidência de perdas de energia eléctrica para depois descobrir os clientes fraudulentos pertencentes a estas localidades aplicando *Data Mining.*

Optou-se pela utilização do *OLAP* para a análise de possíveis perdas nos postos de transformação porque permite visualizar em vários dimensões e de forma rápida e fácil os dados de forma a examinar os mesmos gradualmente. De salientar que não foi possível fazer o balanço energético (diferença entre a energia medida pelos contadores dos postos de transformação e a energia medida pelos contadores dos clientes finais) porque foi utilizada apenas uma amostra e não a totalidade dos clientes da zona seleccionada para descoberta de possíveis perdas de energia eléctrica.

Portanto, as interpretações dos resultados com a utilização de *Data Warehouse* e *OLAP (On-Line Analytical Processing)* para detecção de perdas nos postos de transformação e identificação de zonas críticas com maior incidência de fraudes permitem concluir que os objectivos foram alcançados.

Para o estudo realizado foram identificados alguns indicadores e verificou-se que para descobrir o conhecimento existente é necessário definir algumas regras. Foram definidas três regras que determinaram a forma de classificar um cliente como sendo ou não suspeito de fraudes. Entretanto, pode-se frisar que o uso dessas regras sem a ajuda das ferramentas adequadas trata-se de uma tarefa árdua visto que, envolve diversos cálculos e grandes volumes de dados.

Pode-se concluir que diversos padrões foram identificados por meio da aplicação da descoberta de conhecimento, porém, há muitas outras descobertas que ainda poderiam ser feitas aproveitando-se a *Data Warehouse* criado ou mesmo inserindo novos indicadores.

Muitos são as tarefas e os algoritmos de *Data Mining* que se propõem a extrair conhecimento de bases de dados, e não há uma forma de determinar qual deles é o melhor, visto que os mesmos podem ter melhor desempenho em uma determinada situação e outros podem ser mais eficiente em outros tipos de situações. Neste âmbito optou-se por utilizar a tarefa de classificação e o algoritmo árvore de decisão J48 o que demonstraram ser eficazes para identificação de perdas por serem facilmente explicável, uma vez que tem a forma de regras explícitas.

Observou-se no entanto que, ao gerar as regras de classificação, obtiveram também novos padrões dos dados. A nova informação descoberta pode ser aplicada nas zonas analisadas, de forma a propor medidas correctivas e preventivas para minimizar o problema das possíveis perdas de energia eléctrica.

De uma forma geral, foi possível constatar que a descoberta de conhecimento foi ocorrendo gradualmente, a medida que o processo de *Data Mining* foi sendo concretizado. Na primeira etapa, definição do problema, optou-se por explorar e seleccionar os dados da base de dados referente à rede de distribuição, facturação e consumo do cliente. A segunda etapa de limpeza e transformação dos dados, levou ao contacto mais profundo com os dados extraindo apenas aqueles potencialmente interessantes para a descoberta dos padrões. Na terceira etapa, a realização do *Data Mining* propriamente dito, optou-se pela utilização do software *Weka*, que já traz embutido no seu sistema os algoritmos árvores de decisão J48 e as tarefas de classificação para descoberta do conhecimento. As primeiras análises e constatações advêm desta fase. Na quarta etapa, análise dos dados, novas classificações foram realizadas e o conhecimento emergiu.

Os resultados obtidos mostram que as áreas consideradas críticas da empresa,

continuam sendo as zonas consideradas de classe baixa com maior índice de clientes seleccionados para inspecção com possíveis perdas de energia eléctrica. Mas, de salientar que os clientes que consomem mais energia estão nas zonas consideradas de classe média e alta e neste âmbito, foram analisados 100 clientes de cada uma dessas zonas, os resultados obtidos mostram que a zona de classe alta está com um índice de perdas elevado o que implica inspecções urgentes nesta zona. O sistema de suporte a decisão conseguiu diferenciar correctamente os registos ou valores classificados como normais dos registos classificados com algum grau de suspeita para inspecção.

Deste modo, pode-se concluir que os padrões encontrados podem ser aplicados no terreno para depois serem validados.

Os indicadores obtidos e a sua aplicação reforçam o objectivo do processo de *Data Mining* no sentido de transformar dados em informação a ser utilizada no processo decisório da organização em causa, para a tomada de decisões relacionadas com as perdas de energia eléctrica.

Com a realização deste trabalho, pode-se dizer que atingiu-se os objectivos inicialmente traçados, na medida em que, permitiu-nos aprofundar e consolidar os nossos conhecimentos sobre as tecnologias estudadas, analisando e avaliando os resultados obtidos sobre a detecção de fraudes de energia eléctrica bem como a sua recomendação para a sua utilização na secção comercial da empresa, mostrando o seu grande potencial para o uso imediato.

5.1. Contribuições, dificuldades e desafios

Pelo que se levantou neste estudo, acreditamos que as pesquisas feitas e as aplicações dos conceitos apreendidos sobre o sistema de suporte a decisão em um ambiente real contribuam para o intercâmbio entre profissionais da área em estudo e para o aumento da qualidade e eficiência na gestão de informação para a tomada de decisões por parte dos gestores da empresa.

O desenvolvimento do sistema de suporte de informação à detecção de perdas de energia eléctrica foi essencial pois foi utilizado para modelar um sistema e assim obter conhecimento sobre as possíveis perdas de energia eléctrica e utilizar este conhecimento em prol da selecção de suspeitos de fraudes para inspecções.

De uma forma geral, foi possível constatar que a utilização dessas tecnologias na empresa para descoberta do conhecimento, deverá abrir novas oportunidades de reflexão estratégica em torno das principais dimensões do modelo de negócio. Essas tecnologias só contribuirão para a diferenciação e concretização da criação de valor de forma sustentada quando aplicada de forma inovadora e rentável.

Os problemas que podemos realçar são os mais comuns, principalmente com o arranque do estudo, em saber como fazer, que modelo escolher, para obter um produto final melhor, mas que foi resolvido com o decorrer do tempo.

De salientar um facto importante em relação à demora com o processo de selecção, preparação, tratamento e transformação dos dados, consumindo a maior parte do tempo dedicado ao processo de descoberta do conhecimento, uma vez que os dados encontravam-se fora dos parâmetros normais de descoberta do conhecimento, em forma de códigos em que foi necessária a transformação desses valores em dados significativos.

Um outro problema identificado, o acesso aos vários tipos de dados que é de natureza restrita e sigilosa em que muitos atributos fornecidos estavam codificados para não fornecerem informações claras quanto aos dados que estavam sendo utilizados.

De realçar ainda que, actualmente está decorrer o processo de inspecção de vários clientes em diferentes zonas geográficas, em que se demorou um pouco a espera dos resultados finais destas inspecções para efeitos de comparação dos mesmos. Estes resultados finais acabaram por não sair antes da entrega do trabalho, portanto não houve comparação.

Pode-se acrescentar também uma não assimilação muito concisa do *Software Weka*, mas que foi tornando cada vez mais perceptível e muito aprazível no processo de exploração.

5.2. Perspectivas de continuidade

Continuação com a analise e exploração dos dados para descoberta de novos conhecimentos inserindo novos atributos e parâmetros.

A validação dos dados pelos técnicos da empresa com inspecções no terreno. Após a validação por parte dos técnicos, o estudo poderá ser colocado em prática pela empresa adaptando o mesmo para o efeito, visto que necessita de se trabalhar com uma equipa interdisciplinar, relacionando os conhecimentos dos técnicos com aqueles que dominam as

tecnologias, em especial *Data Mining*.

E por fim, implementar uma interface gráfica que permita a edição de regras numa interface amigável.

6. Bibliografias

Araújo, A. C. e outros, Considerações Sobre as perdas na distribuição de energia eléctrica no Brasil. Trabalho técnico seleccionado para o SENDI 2006.

Araújo, D. L. A. et al. A parallel genetic algorithm for rule Discovery in large databases. IEEE Systems, Man and Cybernetics Conference, Tokyo. v. 3, 1999.

Aurélio, Marco, Marley Vellasco e Carlos Henrique Lopes. Descoberta do conhecimento e mineração de dados. Artigo, ICA _ laboratório de Inteligência Computacional Aplicada, Departamento de Engenharia Eléctrica PUC – Rio, 2000.

Baptista, Gustavo Henrique de Almeida Prado Alves. Pré Processamento de dados em aprendizado de Maquina Supervisionado. Tese de Doutorado, Ciências de Computação e Matemática Computacional, Instituto de Ciências Matemática e Computação – ICMC_USP, São Carlos –SP, 2003.

Berry, Michael J. A.; Gordon Linoff, "Data Mining Techiques for Marketing, and Customer Support"; John Wiley & Sons, Inc., 1997.

Brachman, Ronald J., Tej Arnand, The process of knowledge discovery in databases, advances in knowledge discovery and data mining, American Association for

Artificial Intelligence, Menlo Park, CA, 1996.

Braga, António de Pádua, Teresa Bernarda Ludermir, André Carlos P. de Carvalho; "Redes Neurais Artificiais – Teoria e aplicações ", Editora I 2000.

Brause, R., Langsdorf, T., Hepp, M., Neural Data Mining for Credit Card Fraud Detection, J.W. Goethe-University, 1999.

Burge, P., Taylor, J.S., Moreau, Y., Verrelst, H., Stoermann C, e Gosset, P., Brutus – Hydrid detection Tool, ACTS Mobile Summit 97, Proceedings of ACTS Mobile Summit1997.

Cabral, J. E.; Pinto, J.O.P; Gontijo, E. M.; Reis, J. Rough Sets Based Fraud Detection in Electrical Energy Consumers. WSEAS International Conference on Mathematics And Computers in Phisics, Cancun, Mexico, Apr 2004.

Calili, Rodrigo Flora. Desenvolvimento de sistema para detecção de perdas comerciais em redes de distribuição de energia eléctrica. 2005. 157f. Dissertação (Mestrado em Engenharia Eléctrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2005.

Chen, M. S., J. Han, P. S. Yu, Data Mining: An Overview from Database Perspective, *IEEE Transactions on Knowledge and Data Engineering*, 8(6), Dec 1996.

Corey, M., et al.(Eds), "Oracle Data Warehouse 8i", Oracle Press, 2001. Coutinho, Fernando V., "Data Mining".

Cratochvil, A. *Data Mining* techniques in supporting decision making. Master thesis, Universiteit Leiden, 1999.

Delaiba, A. C. , Reis Filho, J. , Gontijo, E. M. , Mazina, E. , Cabral, J.E. ; Pinto, J. O. P . "Fraud Identification in Electricity Company Costumers Using Decision Tree". In: International Conference on System, Man and Cybernetics, 2004.

Eller, N. A., Arquitectura de informação para gestão de perdas comerciais de energia eléctrica, Programa de Pós Graduação, Engenharia da Produção, UFSC, 2003.

Engels, R. e Theusinger, C. *Using a Data Metric for Preprocessing Advice for Data Mining Applications*, European Conference on Artificial Intelligence, ECAI 1998.

Fayyad U., Piatetsky-Shapiro G. & Smyth P. From Data Mining to Knowledge Discovery: An Overview, in Advances in Knowledge Discovery in Databases, U. Fayyad et al (Eds.) MIT Press. Cambridge, MA, USA, 1996.

Feldman, Ronen & Dagan, Ido. (1995). Knowledge discovery in textual databases.

Freitas, A..A. Understanding the crucial differences between classification and discovery of association rules. A position paper. *SIGKDD Explorations,* 2000.

Freitas, A.A. A genetic algorithm for generalized rule induction. In: WSC3, Onlie Conference On Soft Computing, Hosted On The internet, 3., 1998, Cranfield. Proceedings…Cranfield: Cranfield University, 1999.

Gardner, S.R. Building the Data Warehouse, *Communications of the ACM,* September, (1998).

Goebel, M.; Gruenwald, L. A survey of data mining and knowledge discovery software tools. In: *SIGKDD Explorations*, June 1999.

Goldschmidt, Ronaldo; Passos. Emanuel. Data Mining: Um Guia Prático. Rio de Janeiro: Elsevier, 2005.

Gupta, V. R. An introduction to *Data Warehousing*. System Services Corporation, August 1997.

Han, Jiawei, e Micheline Kamber. Data Mining: Concepts and Techniques. Second edition. Elsevier Science & Tecnology Books, 2006.

Han, Kamber e Kaufmann - *Data Mining: Concepts and Techniques*, J. Han & M. Kamber, Morgan Kaufmann, 2001.

Harrison, T. H., *Intranet Data Warehouse*. McGraw-Hill, USA, 1998.

Haykin, S. Redes neuronais*:* princípios e prática. Porto Alegre: Bookman, 2001.

Inmon, William H. Building the Data Warehouse: Getting Started. 4ªEdição. Editora: Wiley Publishing, inc, 2005.

Kimball, R., *The Data Warehouse Lifecicle Toolkit*. John Wiley & Sons Inc., New York, 1998.

Kimball, Ralph; Ross, Margy. The *Data Warehouse* Toolkit. Guia Completo para Modelação Dimensional. Tradução da 2. ed. Rio de Janeiro: Campus, 2002.

Mannila, H. Methods and problems in *Data Mining*. In*: International Conference on Database Theory*, Delphi, Greece, January 1997.

Minussi, MArlon Mendes. Metodologia de Mineração de Dados Para Detecção de Desvios de Comportamento do Uso de Energia Eléctrica. Dissertação de Mestrado, Programa de Pós Graduação em Engenharia Eléctrica, Pontifícia Universidade Católica do Rio Grande do Sul., 2008.

Passini, S.R.R.; Toledo, Mineração de Dados para Detecção de Fraudes em Ligações de Água. Dissertação de Mestrado. PUC-Campinas. Mar 2002.

PEAS, Programa de Energia, Água e Saneamento – Relatório do programa da Realização de Novas Ligações Domiciliares nas Redes da Electra, Agosto 2002.

Pereira, Celina Maria Rodrigues. Comparação de Ferramentas de *Data Mining*. Monografia, Departamento de Engenharia Informática, Instituto politécnico do Porto, 2002.

Queiroga, Rodrigo Mendonça, 1965-Q3U – Uso de Técnicas de *Data Mining* para Detecção de Fraudes em Energia Eléctrica, Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil, 2005.

Queyroi, Roberto. Aplicação do Modelo de Mineração de dados em um sistema de apoio a decisão para empresas de saneamento. Dissertação de Mestrado. Sistemas Computacionais em Engenharia Civil, Universidade Federal do Rio de Janeiro – UFRJ 2007.

Reis, J. Filho: Gontijo, E.M., Fraude Identification In Electricity Company Costumers Using Decision Tree Systems, Man and Cybernetics, 2004.

Rezende, S.O., Pugliese, J. B. e Varejão, F.M. Sistemas Baseados em Conhecimentos, Sistemas Inteligentes, capitulo2, Rezende, S.O (coordenadora), Editora Manole, 2003.

Sanches, André Rodrigo. Uma visão geral sobre *Data Mining*. Relatório de Estudo – Tópicos em Ciência da Computação, Dept. Ciência da Computação, Universidade de São Paulo – USP, São Paulo, 2003.

Santos, Ricardo da Silva. Ambiente para extracção de informações através da Mineração de Dados do Sistema Único de Saúde. Tese de Doutorado, Programa de Pós Graduação em informática em saúde, Universidade Federal de São Paulo – USP, São Paulo, 2007.

Sarawagi, S.; Agrawal, R.; Megiddo, N. "Discovery-Driven Exploration of OLAP Data Cubes". IBM Almaden Research Center, 1998.

Sferra e Corrêa, Heloisa Helena Sferra, Ângela M. C. Jorge Corrêa. Conceitos e Aplicações de Data Mining, 2003.

Shafer J. & Agrrawal R.. " Parallel Mining of Association Rules". IEEE Trans. Knowledge and Data Eng., Vol. 8. No. 6. pp 19-30, Dec. 1996.

Silva, M. P. S.; Robin, J. R. "SKDQL – Uma Linguagem Declarativa de Especificação de Consultas e Processos para Descoberta de Conhecimento em Bancos de Dados e sua Implementação" (2002). Dissertação de Mestrado. UFPE, 2003.

Silva, M.P. (November de 2004). Mineração de Dados – Conceitos, aplicações e experiência com Weka. Livro da escola regional de Informática de Rio de Janeiro – Espírito Santo. Rio Janeiro: SBC.

Simoff, S.; Djeraba, C.; Zaiane, O. "Multimedia *Data Mining* between Promisses and Problems" (2002). SIGKDD Explorations.

Viveros, M.S.; Nearhos, J.P.; Rothman, M.J. Applying data mining techniques to a health insurance information system. In: *22nd VLDB Conference*, Mumba (Bombay), India, 1996.

Wang, John Wang, Data Warehousing and Mining: Conceps, Methodologies, Tools, and Applications, Montclair State University, 2008.

Weiss, S. & Kulikowski, C. *Computer Systems that Learn*. San Mateo, CA: Morgan Kaufmann, 1991

Yao, Y.Y. et al. (2001). Web Intelligence (WI): research challenges and trends in the new

Sites utilizados para consultas

Anatel: http://www.anatel.gov.br/ - Acedida em Maio de 2009

Aneel: http://www.aneel.gov.br/ - Acedida em Maio de 2009

Escelsa: http://www.escelsa.com.br/ - Acedida em Junho de 2009

Ampla: http://www.ampla.com/ - Acedida em Junho de 2009

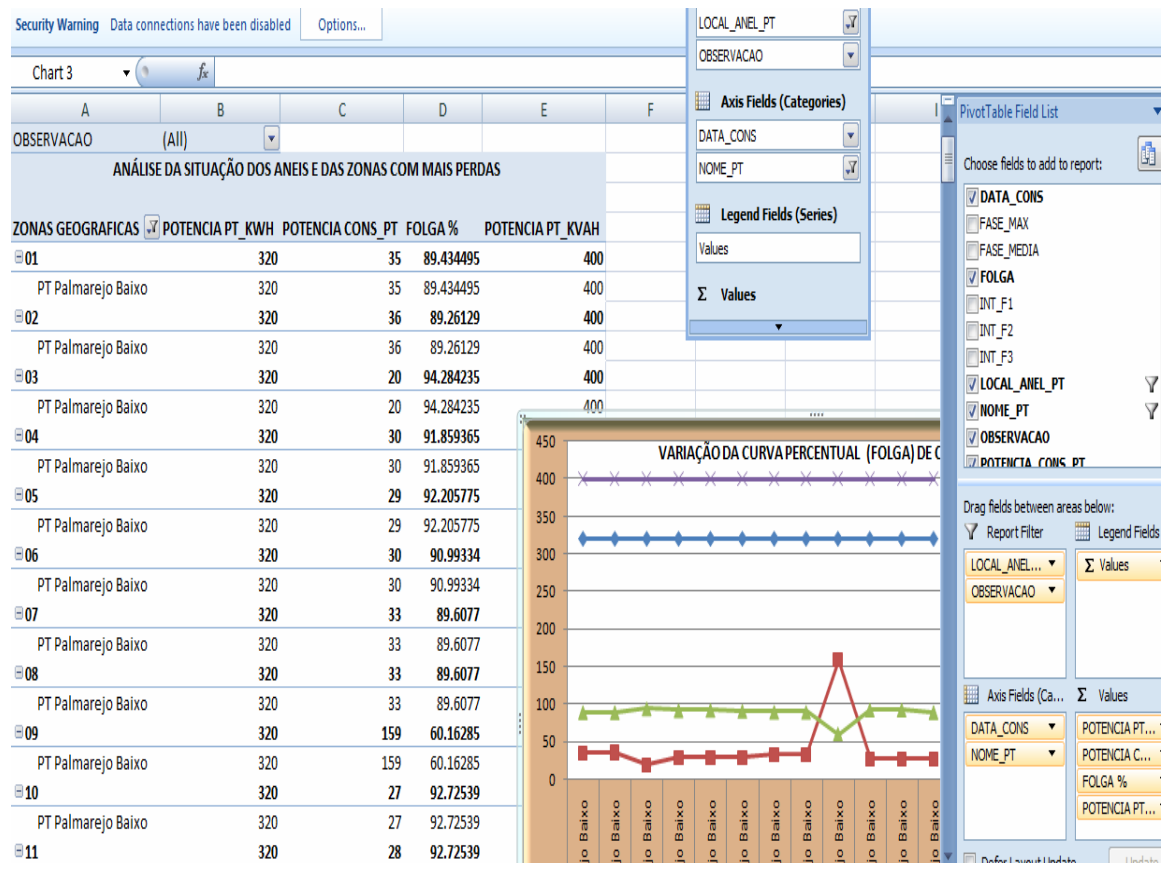
INE: Instituto Nacional de Estatística – http://www.ine.cv/

Electra: Empresa de Produção e Distribuição de Energia e Água: http://www.electra.cv/

Anexos

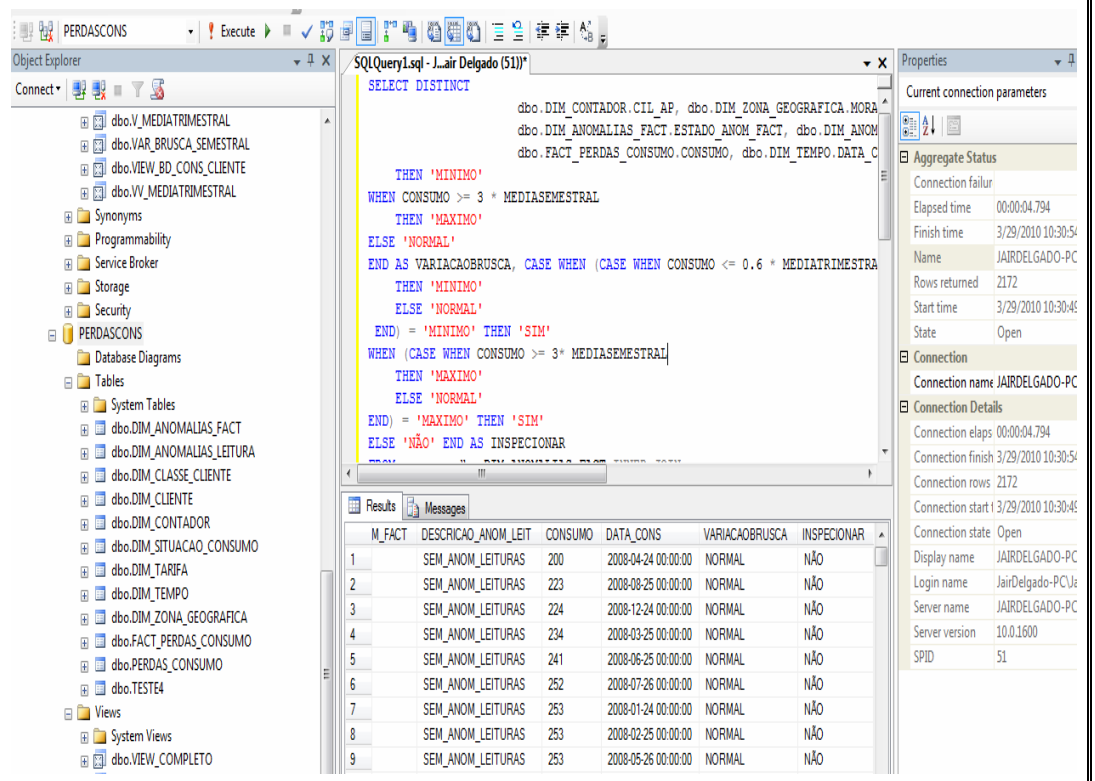
Análise dos PT´s com OLAP

ANÁLISE DOS POSTOS DE TRANSFORMAÇÃO COM OLAP



Criação de View´s na Base de Dados

CRIACÇÃO DAS VIEW'S NA BASE DE DADOS

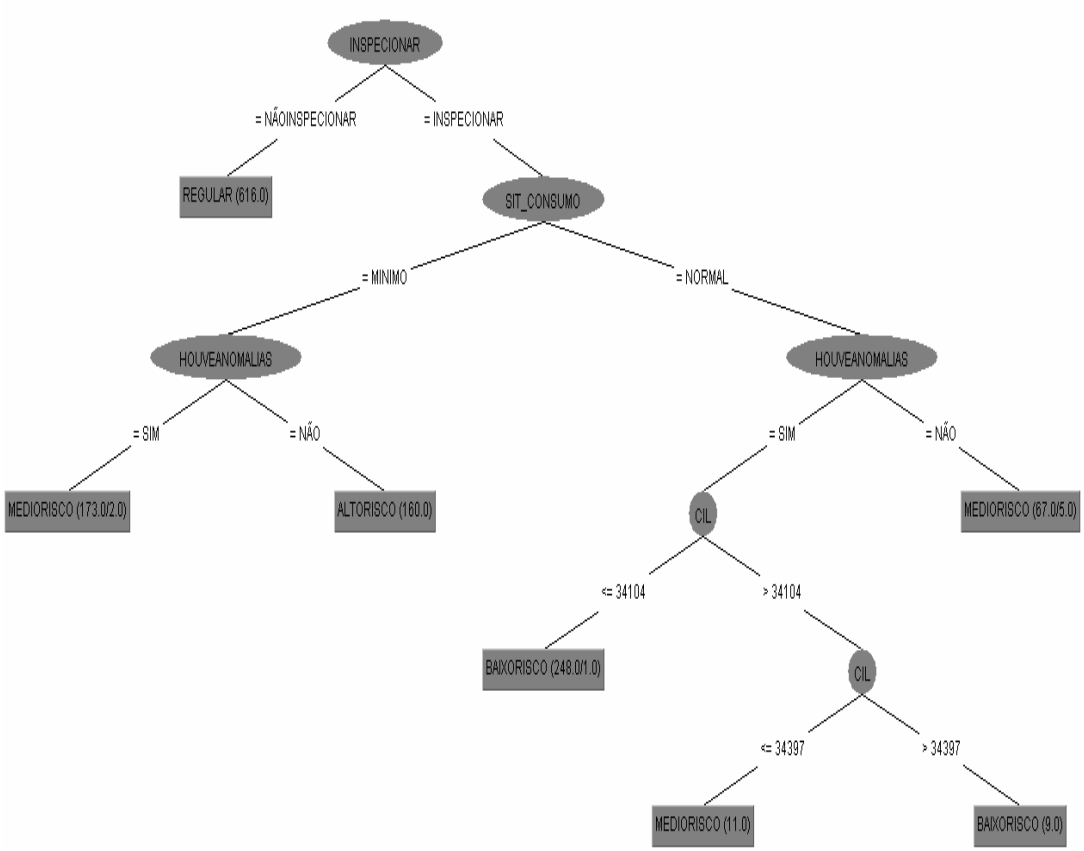


Análise dos resultados com *Weka*

SUSPEITAS DE PERDAS – ZONA DE CLASSE BAIXA

```
=== Run information ===
Relation:     SUSPEITASPERDASPENSAMENTO
Instances:    1284
Attributes:   11
              CIL
              MARCA
              TP_CLIENTE
              TIPO_FACT
              DESCRICAO_ANOM_FACT
              ESTADOANOM_FACT
              TARIFA
              SIT_CONSUMO
              HOUVEANOMALIAS
              SUSPEITASPERDAS
              INSPECIONAR

=== Classifier model (full training set) ===

INSPECIONAR = NÃOINSPECIONAR: REGULAR (616.0)
INSPECIONAR = INSPECIONAR
|   SIT_CONSUMO = MINIMO
|   |   HOUVEANOMALIAS = SIM: MEDIORISCO (173.0/2.0)
|   |   HOUVEANOMALIAS = NÃO: ALTORISCO (160.0)
|   SIT_CONSUMO = NORMAL
|   |   HOUVEANOMALIAS = SIM
|   |   |   CIL <= 34104: BAIXORISCO (248.0/1.0)
|   |   |   CIL > 34104
|   |   |   |   CIL <= 34397: MEDIORISCO (11.0)
|   |   |   |   CIL > 34397: BAIXORISCO (9.0)
|   |   HOUVEANOMALIAS = NÃO: MEDIORISCO (67.0/5.0)

Number of Leaves  : 7
Time taken to build model: 0.07 seconds
=== Summary ===

Correctly Classified Instances        1276               99.3769 %
Incorrectly Classified Instances         8                0.6231 %
Kappa statistic                          0.9908
Mean absolute error                      0.0059
Root mean squared error                  0.0544
```

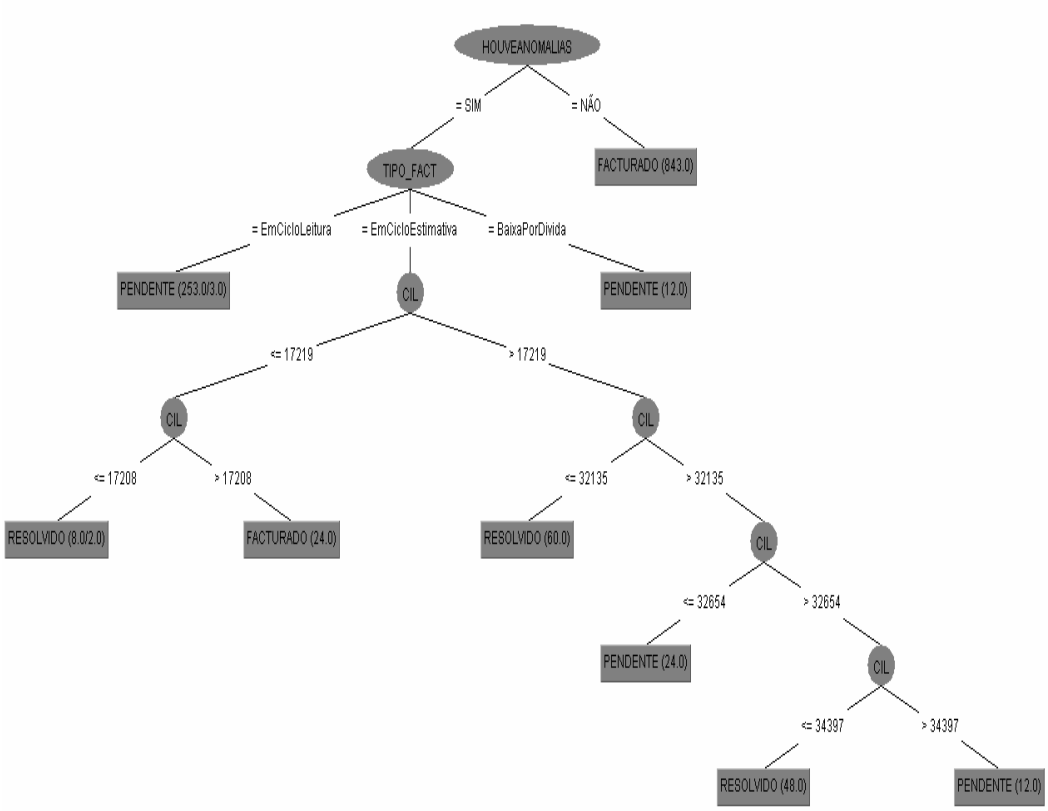
```
Relative absolute error               1.7517 %
Root relative squared error           13.2373 %
Total Number of Instances             1284

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.994    0          1          0.994     0.997      1         ALTORISCO
                 0.996    0.007      0.972      0.996     0.984      0.998     MEDIORISCO
                 0.977    0.001      0.996      0.977     0.987      0.998     BAIXORISCO
                 1        0          1          1         1          1         REGULAR
Weighted Avg.    0.994    0.001      0.994      0.994     0.994      0.999

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 160   1   0   0 |   a = ALTORISCO
   0 244   1   0 |   b = MEDIORISCO
   0   6 256   0 |   c = BAIXORISCO
   0   0   0 616 |   d = REGULAR
```



FACTURAÇÃO DOS CLIENTES QTO A ANOMALIAS – ZONA DE CLASSE BAIXA

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     FACTURAÇÃODOSCLIENTES
Instances:    1284
Attributes:   11
              CIL
              MARCA
              TP_CLIENTE
              TIPO_FACT
              DESCRICAO_ANOM_FACT
              ESTADOANOM_FACT
              TARIFA
              SIT_CONSUMO
              HOUVEANOMALIAS
              SUSPEITASPERDAS
              INSPECIONAR
Test mode:    evaluate on training data
=== Classifier model (full training set) ===

HOUVEANOMALIAS = SIM
|   TIPO_FACT = EmCicloLeitura: PENDENTE (253.0/3.0)
|   TIPO_FACT = EmCicloEstimativa
|   |   CIL <= 17219
|   |   |   CIL <= 17208: RESOLVIDO (8.0/2.0)
|   |   |   CIL > 17208: FACTURADO (24.0)
|   |   CIL > 17219
|   |   |   CIL <= 32135: RESOLVIDO (60.0)
|   |   |   CIL > 32135
|   |   |   |   CIL <= 32654: PENDENTE (24.0)
|   |   |   |   CIL > 32654
|   |   |   |   |   CIL <= 34397: RESOLVIDO (48.0)
|   |   |   |   |   CIL > 34397: PENDENTE (12.0)
|   TIPO_FACT = BaixaPorDivida: PENDENTE (12.0)
HOUVEANOMALIAS = NÃO: FACTURADO (843.0)

Number of Leaves  :  9

Size of the tree :        16
Time taken to build model: 0.01 seconds
=== Summary ===

Correctly Classified Instances        1279               99.6106 %
Incorrectly Classified Instances         5                0.3894 %
Kappa statistic                          0.9919
```

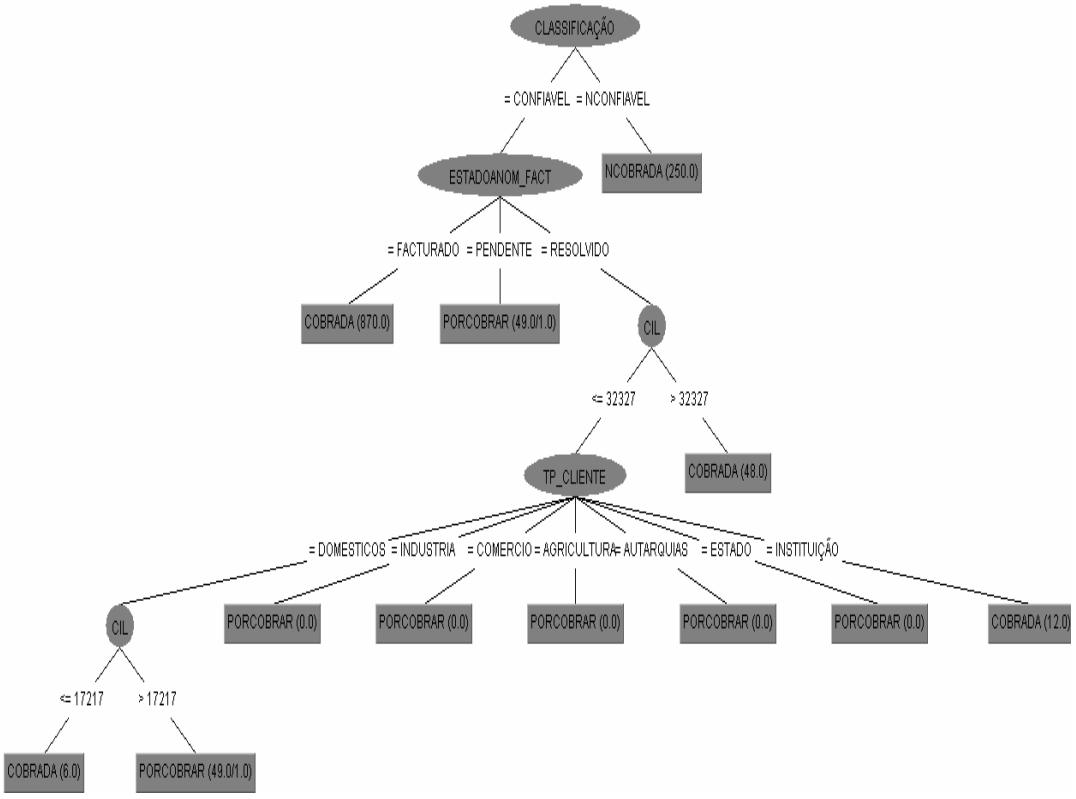
```
Mean absolute error                0.0046
Root mean squared error            0.0482
Relative absolute error            1.4562 %
Root relative squared error        12.0727 %
Total Number of Instances          1284

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.995    0          1         0.995     0.998      0.999    FACTURADO
                 1        0.003      0.99      1         0.995      0.999    PENDENTE
                 0.991    0.002      0.983     0.991     0.987      0.999    RESOLVIDO
Weighted Avg.    0.996    0.001      0.996     0.996     0.996      0.999

=== Confusion Matrix ===
   a   b   c   <-- classified as
 867   2   2 |   a = FACTURADO
   0 298   0 |   b = PENDENTE
   0   1 114 |   c = RESOLVIDO
```



CONFIABILIDADE DOS CLIENTES – ZONA DE CLASSE BAIXA

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     CONFIABILIDADEPENSAMENTO
Instances:    1284
Attributes:   10
              CIL
              CONSUMO
              VALOR_FACT
              TP_CLIENTE
              DESCRICAO_ANOM_FACT
              TIPO_FACT
              ESTADOANOM_FACT
              TARIFA
              ESTADOFACT
              CLASSIFICAÇÃO
Test mode:    evaluate on training data
J48 pruned tree
------------------

CLASSIFICAÇÃO = CONFIAVEL
|   ESTADOANOM_FACT = FACTURADO: COBRADA (870.0)
|   ESTADOANOM_FACT = PENDENTE: PORCOBRAR (49.0/1.0)
|   ESTADOANOM_FACT = RESOLVIDO
|   |   CIL <= 32327
|   |   |   TP_CLIENTE = DOMESTICOS
|   |   |   |   CIL <= 17217: COBRADA (6.0)
|   |   |   |   CIL > 17217: PORCOBRAR (49.0/1.0)
|   |   |   TP_CLIENTE = INDUSTRIA: PORCOBRAR (0.0)
|   |   |   TP_CLIENTE = COMERCIO: PORCOBRAR (0.0)
|   |   |   TP_CLIENTE = AGRICULTURA: PORCOBRAR (0.0)
|   |   |   TP_CLIENTE = AUTARQUIAS: PORCOBRAR (0.0)
|   |   |   TP_CLIENTE = ESTADO: PORCOBRAR (0.0)
|   |   |   TP_CLIENTE = INSTITUIÇÃO: COBRADA (12.0)
|   |   CIL > 32327: COBRADA (48.0)
CLASSIFICAÇÃO = NCONFIAVEL: NCOBRADA (250.0)
Number of Leaves  :  12
Size of the tree :        17
Time taken to build model: 0.07 seconds

Correctly Classified Instances        1282               99.8442 %
Incorrectly Classified Instances         2                0.1558 %
Kappa statistic                          0.9963
```

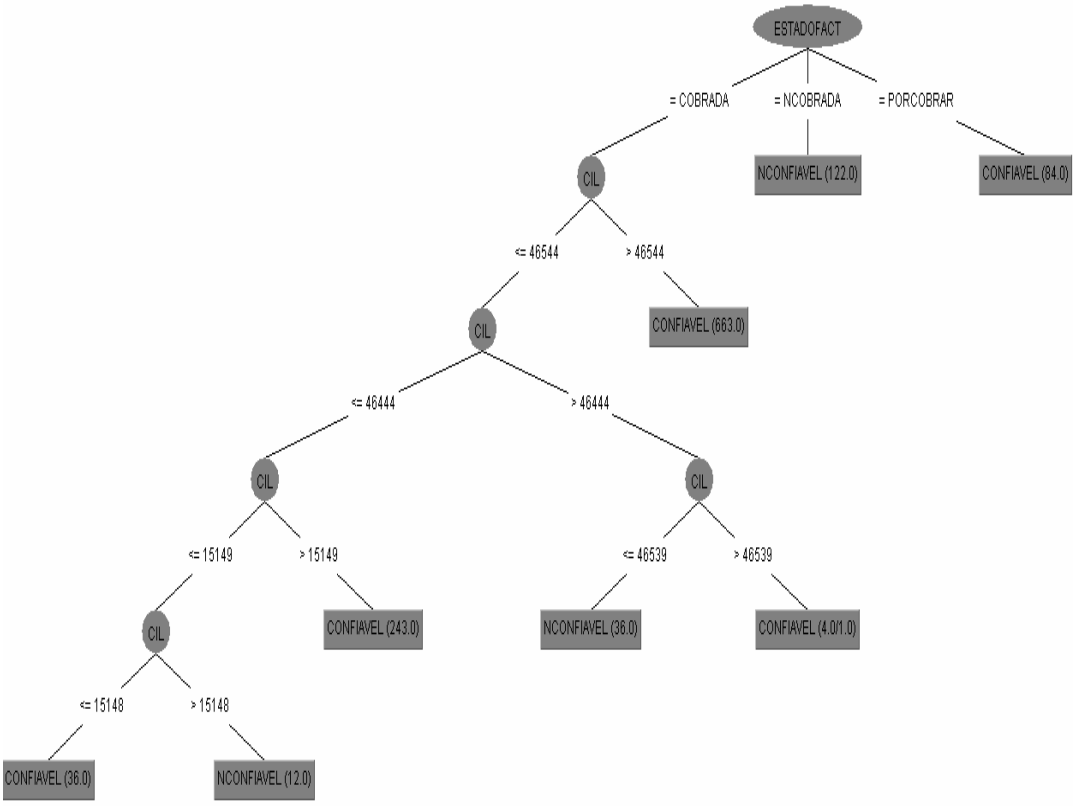
```
K&B Relative Info Score          127365.1523 %
K&B Information Score             1368.3503 bits      1.0657 bits/instance
Class complexity | order 0        1376.1525 bits      1.0718 bits/instance
Class complexity | scheme           14.0852 bits      0.011  bits/instance
Complexity improvement     (Sf)    1362.0673 bits      1.0608 bits/instance
Mean absolute error               0.002
Root mean squared error             0.0319
Relative absolute error           0.7194 %
Root relative squared error         8.4871 %
Total Number of Instances         1284

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     0          1        0.999     0.999      1        COBRADA
                 0.996     0          1        0.996     0.998      1        NCOBRADA
                 1         0.002      0.98     1         0.99       0.999    PORCOBRAR
Weighted Avg.    0.998     0          0.998    0.998     0.998      1

=== Confusion Matrix ===

   a   b   c   <-- classified as
 936   0   1 |   a = COBRADA
   0 250   1 |   b = NCOBRADA
   0   0  96 |   c = PORCOBRAR
```



CONFIABILIDADE DOS CLIENTES – ZONA DE CLASSE ALTA

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     CONFIABILIDADEPALMAREJO
Instances:    1200
Attributes:   10
              CIL
              CONSUMO
              VALOR_FACT
              TP_CLIENTE
              DESCRICAO_ANOM_FACT
              TIPO_FACT
              ESTADOANOM_FACT
              TARIFA
              ESTADOFACT
              CLASSIFICAÇÃO
Test mode:    evaluate on training data
J48 pruned tree
------------------

ESTADOFACT = COBRADA
|   CIL <= 46544
|   |   CIL <= 46444
|   |   |   CIL <= 15149
|   |   |   |   CIL <= 15148: CONFIAVEL (36.0)
|   |   |   |   CIL > 15148: NCONFIAVEL (12.0)
|   |   |   CIL > 15149: CONFIAVEL (243.0)
|   |   CIL > 46444
|   |   |   CIL <= 46539: NCONFIAVEL (36.0)
|   |   |   CIL > 46539: CONFIAVEL (4.0/1.0)
|   CIL > 46544: CONFIAVEL (663.0)
ESTADOFACT = NCOBRADA: NCONFIAVEL (122.0)
ESTADOFACT = PORCOBRAR: CONFIAVEL (84.0)

Number of Leaves  :         8
Size of the tree :    14
Time taken to build model: 0.12 seconds
```

```
=== Summary ===

Correctly Classified Instances        1199               99.9167 %
Incorrectly Classified Instances         1                0.0833 %
Kappa statistic                          0.9966
K&B Relative Info Score             118830.563  %
K&B Information Score                 703.8178 bits      0.5865 bits/instance
Class complexity | order 0            708.9011 bits      0.5908 bits/instance
Class complexity | scheme               3.2451 bits      0.0027 bits/instance
Complexity improvement     (Sf)       705.656  bits      0.588  bits/instance
Mean absolute error                      0.0013
Root mean squared error                  0.025
Relative absolute error                  0.5106 %
Root relative squared error              7.1518 %
Total Number of Instances             1200
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.006      0.999     1         1          1        CONFIAVEL
                 0.994     0          1         0.994     0.997      1        NCONFIAVEL
Weighted Avg.    0.999     0.005      0.999     0.999     0.999      1

=== Confusion Matrix ===

   a    b   <-- classified as
1029    0 |    a = CONFIAVEL
   1  170 |    b = NCONFIAVEL
```



VARIAÇÃO BRUSCA – ZONA DE CLASSE ALTA

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: VARIAÇÃOBRUSCA
Instances: 1200
Attributes: 10
CIL
TP_CLIENTE
CONSUMO
VALOR_FACT
TIPO_FACT
ESTADOFACT
HOUVEANOMALIAS
VARIAÇÃOBRUSCA
SUSPEITASPERDAS
INSPECIONAR

```
Test mode:    evaluate on training data
J48 pruned tree
------------------

SUSPEITASPERDAS = ALTORISCO: INSPECIONAR (192.0)
SUSPEITASPERDAS = MEDIORISCO: INSPECIONAR (232.0/13.0)
SUSPEITASPERDAS = BAIXORISCO
|   TIPO_FACT = EmCicloLeitura: INSPECIONAR (85.0)
|   TIPO_FACT = EmCicloEstimativa
|   |   VARIAÇÃOBRUSCA = MINIMO: INSPECIONAR (0.0)
|   |   VARIAÇÃOBRUSCA = MAXIMO
|   |   |   CIL <= 56400: NÃOINSPECIONAR (10.0)
|   |   |   CIL > 56400: INSPECIONAR (20.0)
|   |   VARIAÇÃOBRUSCA = NORMAL: INSPECIONAR (21.0)
|   TIPO_FACT = BaixaporDívida: INSPECIONAR (0.0)
SUSPEITASPERDAS = REGULAR: NÃOINSPECIONAR (640.0/1.0)

Number of Leaves  :         9

Size of the tree :     13
Time taken to build model: 0.1 seconds

=== Evaluation on training set ===

Correctly Classified Instances        1186               98.8333 %
Incorrectly Classified Instances        14                1.1667 %
Kappa statistic                          0.9765
K&B Relative Info Score             115281.2929 %
K&B Information Score                 1143.9473 bits      0.9533 bits/instance
Class complexity | order 0            1190.7406 bits      0.9923 bits/instance
Class complexity | scheme               83.031  bits      0.0692 bits/instance
Complexity improvement     (Sf)       1107.7096 bits      0.9231 bits/instance
Mean absolute error                      0.0221
Root mean squared error                  0.1052
Relative absolute error                  4.471  %
Root relative squared error             21.1449 %
Total Number of Instances             1200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.98      0.002      0.998     0.98      0.989      0.995    NÃOINSPECIONAR
                 0.998     0.02       0.976     0.998     0.987      0.995    INSPECIONAR
Weighted Avg.    0.988     0.01       0.989     0.988     0.988      0.995

=== Confusion Matrix ===
```

```
  a   b   <-- classified as
649  13 |   a = NÃOINSPECIONAR
  1 537 |   b = INSPECIONAR
```